iclr 2022

Domino: Discovering Systematic Errors with Cross-Modal Embeddings

- Sabri Eyuboglu, Maya Varma, Khaled Kamal Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, Christopher Re
- abstract@open-review(Oral): Machine learning models that achieve high overall accuracy often make systematic errors on important subsets (or slices) of data. Identifying underperforming slices is particularly challenging when working with high-dimensional inputs (e.g. images, audio), where important slices are often unlabeled. In order to address this issue, recent studies have proposed automated slice discovery methods (SDMs), which leverage learned model representations to mine input data for slices on which a model performs poorly. To be useful to a practitioner, these methods must identify slices that are both underperforming and coherent (i.e. united by a human-understandable concept). However, no quantitative evaluation framework currently exists for rigorously assessing SDMs with respect to these criteria. Additionally, prior qualitative evaluations have shown that SDMs often identify slices that are incoherent. In this work, we address these challenges by first designing a principled evaluation framework that enables a quantitative comparison of SDMs across 1,235 slice discovery settings in three input domains (natural images, medical images, and time-series data). Then, motivated by the recent development of powerful cross-modal representation learning approaches, we present Domino, an SDM that leverages cross-modal embeddings and a novel error-aware mixture model to discover and describe coherent slices. We find that Domino accurately identifies 36% of the 1,235 slices in our framework -- a 12 percentage point improvement over prior methods. Further, Domino is the first SDM that can provide natural language descriptions of identified slices, correctly generating the exact name of the slice in 35% of settings.

Natural Language Descriptions of Deep Visual Features

- Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, Jacob Andreas
- abstract@open-review(Oral): Some neurons in deep networks specialize in recognizing highly specific perceptual, structural, or semantic features of inputs. In computer vision, techniques exist for identifying neurons that respond to individual concept categories like colors, textures, and object classes. But these techniques are limited in scope, labeling only a small subset of neurons and behaviors in any network. Is a richer characterization of neuron-level computation possible? We introduce a procedure (called MILAN, for mutual information-guided linguistic annotation of neurons) that automatically labels neurons with open-ended, compositional, natural language descriptions. Given a neuron, MILAN generates a description by searching for a natural language string that maximizes pointwise mutual information with the image regions in which the neuron is active. MILAN produces fine-grained descriptions that capture categorical, relational, and logical structure in learned features. These descriptions obtain high agreement with human-generated feature descriptions across a diverse set of model architectures and tasks, and can aid in understanding and controlling learned models. We highlight three applications of natural language neuron descriptions. First, we use MILAN for analysis, characterizing the distribution and importance of neurons selective for attribute, category, and relational information in vision models. Second, we use MILAN for auditing, surfacing neurons sensitive to human faces in datasets designed to obscure them. Finally, we use MILAN for editing, improving robustness in an image classifier by deleting neurons sensitive to text features spuriously correlated with class labels.

Non-Transferable Learning: A New Approach for Model Ownership Verification and Applicability Authorization

- Lixu Wang, Shichao Xu, Ruiqi Xu, Xiao Wang, Qi Zhu
- abstract@open-review(Oral): As Artificial Intelligence as a Service gains popularity, protecting well-trained models as intellectual property is becoming increasingly important. There are two common types of protection methods: ownership verification and usage authorization. In this paper, we propose Non-Transferable Learning (NTL), a novel approach that captures the exclusive data representation in the learned model and restricts the model generalization ability to certain domains. This approach provides effective solutions to both model verification and authorization. Specifically: 1) For ownership verification, watermarking techniques are commonly used but are often vulnerable to sophisticated watermark removal methods. By comparison, our NTL-based ownership verification provides robust resistance to state-of-the-art watermark removal methods, as shown in extensive experiments with 6 removal approaches over the digits, CIFAR10 & STL10, and VisDA datasets. 2) For usage authorization, prior solutions focus on authorizing specific users to access the model, but authorized users can still apply the model to any data without restriction. Our NTL-based authorization approach instead provides data-centric protection, which we call applicability authorization, by significantly degrading the performance of the model on unauthorized data. Its effectiveness is also shown through experiments on aforementioned datasets.

Neural Structured Prediction for Inductive Node Classification

- Meng Qu, Huiyu Cai, Jian Tang
- abstract@open-review(Oral): This paper studies node classification in the inductive setting, i.e., aiming to learn a model on labeled training graphs and generalize it to infer node labels on unlabeled test graphs. This problem has been extensively studied with graph neural networks (GNNs) by learning effective node representations, as well as traditional structured prediction methods for modeling the structured output of node labels, e.g., conditional random fields (CRFs). In this paper, we present a new approach called the Structured Proxy Network (SPN), which combines the advantages of both worlds. SPN defines flexible potential functions of CRFs with GNNs. However, learning such a model is nontrivial as it involves optimizing a maximin game with high-cost inference. Inspired by the underlying connection between joint and marginal distributions defined by Markov networks, we propose to solve an approximate version of the optimization problem as a proxy, which yields a near-optimal solution, making learning more efficient. Extensive experiments on two settings show that our approach outperforms many competitive baselines.

A New Perspective on "How Graph Neural Networks Go Beyond Weisfeiler-Lehman?"

- Asiri Wijesinghe, Qing Wang
- abstract@open-review(Oral): We propose a new perspective on designing powerful Graph Neural Networks (GNNs). In a nutshell, this enables a general solution to inject structural properties of graphs into a message-passing aggregation scheme of GNNs. As a theoretical basis, we develop a new hierarchy of local isomorphism on neighborhood subgraphs. Then, we theoretically characterize how message-passing GNNs can be designed to be more expressive than the Weisfeiler Lehman test. To elaborate this characterization, we propose a novel neural model, called GraphSNN, and prove that this model is strictly more expressive than the Weisfeiler Lehman test in distinguishing graph structures. We empirically verify the strength of our model on different graph learning tasks. It is shown that our model consistently improves the state-of-the-art methods on the benchmark tasks without sacrificing computational simplicity and efficiency.

Minibatch vs Local SGD with Shuffling: Tight Convergence Bounds and Beyond

- Chulhee Yun, Shashank Rajput, Suvrit Sra
- abstract@open-review(Oral): In distributed learning, local SGD (also known as federated averaging) and its simple baseline minibatch SGD are widely studied optimization methods. Most existing analyses of these methods assume independent and unbiased gradient estimates obtained via with-replacement sampling. In contrast, we study shuffling-based variants: minibatch and local Random Reshuffling, which draw stochastic gradients without replacement and are thus closer to practice. For smooth functions satisfying the Polyak-Å ojasiewicz condition, we obtain convergence bounds (in the large epoch regime) which show that these shuffling-based variants converge faster than their with-replacement counterparts. Moreover, we prove

matching lower bounds showing that our convergence analysis is tight. Finally, we propose an algorithmic modification called synchronized shuffling that leads to convergence rates faster than our lower bounds in near-homogeneous settings.

The Hidden Convex Optimization Landscape of Regularized Two-Layer ReLU Networks: an Exact Characterization of Optimal Solutions

- Yifei Wang, Jonathan Lacotte, Mert Pilanci
- abstract@open-review(Oral): We prove that finding all globally optimal two-layer ReLU neural networks can be performed by solving a convex optimization program with cone constraints. Our analysis is novel, characterizes all optimal solutions, and does not leverage duality-based analysis which was recently used to lift neural network training into convex spaces. Given the set of solutions of our convex optimization program, we show how to construct exactly the entire set of optimal neural networks. We provide a detailed characterization of this optimal set and its invariant transformations. As additional consequences of our convex perspective, (i) we establish that Clarke stationary points found by stochastic gradient descent correspond to the global optimum of a subsampled convex problem (ii) we provide a polynomial-time algorithm for checking if a neural network is a global minimum of the training loss (iii) we provide an explicit construction of a continuous path between any neural network and the global minimum of its sublevel set and (iv) characterize the minimal size of the hidden layer so that the neural network optimization landscape has no spurious valleys. Overall, we provide a rich framework for studying the landscape of neural network training loss through convexity.

Provably Filtering Exogenous Distractors using Multistep Inverse Dynamics

- Yonathan Efroni, Dipendra Misra, Akshay Krishnamurthy, Alekh Agarwal, John Langford
- abstract@open-review(Oral): Many real-world applications of reinforcement learning (RL) require the agent to deal with high-dimensional observations such as those generated from a megapixel camera. Prior work has addressed such problems with representation learning, through which the agent can provably extract endogenous, latent state information from raw observations and subsequently plan efficiently. However, such approaches can fail in the presence of temporally correlated noise in the observations, a phenomenon that is common in practice. We initiate the formal study of latent state discovery in the presence of such exogenous noise sources by proposing a new model, the Exogenous Block MDP (EX-BMDP), for rich observation RL. We start by establishing several negative results, by highlighting failure cases of prior representation learning based approaches. Then, we introduce the Predictive Path Elimination (PPE) algorithm, that learns a generalization of inverse dynamics and is provably sample and computationally efficient in EX-BMDPs when the endogenous state dynamics are near deterministic. The sample complexity of PPE depends polynomially on the size of the latent endogenous state space while not directly depending on the size of the observation space, nor the exogenous state space. We provide experiments on challenging exploration problems which show that our approach works empirically.

Bootstrapped Meta-Learning

- Sebastian Flennerhag, Yannick Schroecker, Tom Zahavy, Hado van Hasselt, David Silver, Satinder Singh
- abstract@open-review(Oral): Meta-learning empowers artificial intelligence to increase its efficiency by learning how to learn. Unlocking this potential involves overcoming a challenging meta-optimisation problem. We propose an algorithm that tackles this problem by letting the meta-learner teach itself. The algorithm first bootstraps a target from the meta-learner, then optimises the meta-learner by minimising the distance to that target under a chosen (pseudo-)metric. Focusing on meta-learning with gradients, we establish conditions that guarantee performance improvements and show that metric can be used to control meta-optimisation. Meanwhile, the bootstrapping mechanism can extend the effective meta-learning horizon without requiring backpropagation through all updates. We achieve a new state-of-the art for model-free agents on the Atari ALE benchmark and demonstrate that it yields both performance and efficiency gains in multi-task meta-learning. Finally, we explore how bootstrapping opens up new possibilities and find that it can meta-learn efficient exploration in an epsilon-greedy Q-learning agent - without backpropagating through the update rule.

Coordination Among Neural Modules Through a Shared Global Workspace

- Anirudh Goyal, Aniket Rajiv Didolkar, Alex Lamb, Kartikeya Badola, Nan Rosemary Ke, Nasim Rahaman, Jonathan Binas, Charles Blundell, Michael Curtis Mozer, Yoshua Bengio
- abstract@open-review(Oral): Deep learning has seen a movement away from representing examples with a monolithic hidden state towards a richly structured state. For example, Transformers segment by position, and object-centric architectures decompose images into entities. In all these architectures, interactions between different elements are modeled via pairwise interactions: Transformers make use of self-attention to incorporate information from other positions and object-centric architectures make use of graph neural networks to model interactions among entities. We consider how to improve on pairwise interactions in terms of global coordination and a coherent, integrated representation that can be used for downstream tasks. In cognitive science, a global workspace architecture has been proposed in which functionally specialized components share information through a common, bandwidth-limited communication channel. We explore the use of such a communication channel in the context of deep learning for modeling the structure of complex environments. The proposed method includes a shared workspace through which communication among different specialist modules takes place but due to limits on the communication bandwidth, specialist modules must compete for access. We show that capacity limitations have a rational basis in that (1) they encourage specialization and compositionality and (2) they facilitate the synchronization of otherwise independent specialists.

Data-Efficient Graph Grammar Learning for Molecular Generation

- Minghao Guo, Veronika Thost, Beichen Li, Payel Das, Jie Chen, Wojciech Matusik
- abstract@open-review(Oral): The problem of molecular generation has received significant attention recently. Existing methods are typically based on deep neural networks and require training on large datasets with tens of thousands of samples. In practice, however, the size of class-specific chemical datasets is usually limited (e.g., dozens of samples) due to labor-intensive experimentation and data collection. Another major challenge is to generate only physically synthesizable molecules. This is a non-trivial task for neural network-based generative models since the relevant chemical knowledge can only be extracted and generalized from the limited training data. In this work, we propose a data-efficient generative model that can be learned from datasets with orders of magnitude smaller sizes than common benchmarks. At the heart of this method is a learnable graph grammar that generates molecules from a sequence of production rules. Without any human assistance, these production rules are automatically constructed from training data. Furthermore, additional chemical knowledge can be incorporated into the model by further grammar optimization. Our learned graph grammar yields state-of-the-art results on generating high-quality molecules for three monomer datasets that contain only ${\sim}20$ samples each. Our approach also achieves remarkable performance in a challenging polymer generation task with $only$ 117 training samples and is competitive against existing methods using 81k data points.

Poisoning and Backdooring Contrastive Learning

- Nicholas Carlini, Andreas Terzis
- abstract@open-review(Oral): Multimodal contrastive learning methods like CLIP train on noisy and uncurated training datasets. This is cheaper than labeling datasets manually, and even improves out-of-distribution robustness. We show that this practice makes backdoor and poisoning attacks a significant threat. By poisoning just 0.01% of a dataset (e.g., just 300 images of the 3 million-example Conceptual Captions dataset), we can cause the model to misclassify test images by overlaying a small patch. Targeted poisoning attacks, whereby the model misclassifies a particular test input with an

adversarially-desired label, are even easier requiring control of 0.0001% of the dataset (e.g., just three out of the 3 million images). Our attacks call into question whether training on noisy and uncurated Internet scrapes is desirable.

Neural Collapse Under MSE Loss: Proximity to and Dynamics on the Central Path

- X.Y. Han, Vardan Papyan, David L. Donoho
- abstract@open-review(Oral): The recently discovered Neural Collapse (NC) phenomenon occurs pervasively in today's deep net training paradigm of driving cross-entropy (CE) loss towards zero. During NC, last-layer features collapse to their class-means, both classifiers and class-means collapse to the same Simplex Equiangular Tight Frame, and classifier behavior collapses to the nearest-class-mean decision rule. Recent works demonstrated that deep nets trained with mean squared error (MSE) loss perform comparably to those trained with CE. As a preliminary, we empirically establish that NC emerges in such MSE-trained deep nets as well through experiments on three canonical networks and five benchmark datasets. We provide, in a Google Colab notebook, PyTorch code for reproducing MSE-NC and CE-NC: https://colab.research.google.com/github/neuralcollapse/neuralcollapse/blob/main/neuralcollapse.ipynb. The analytically-tractable MSE loss offers more mathematical opportunities than the hard-to-analyze CE loss, inspiring us to leverage MSE loss towards the theoretical investigation of NC. We develop three main contributions: (I) We show a new decomposition of the MSE loss into (A) terms directly interpretable through the lens of NC and which assume the last-layer classifier is exactly the least-squares classifier; and (B) a term capturing the deviation from this least-squares classifier. (II) We exhibit experiments on canonical datasets and networks demonstrating that term-(B) is negligible during training. This motivates us to introduce a new theoretical construct: the central path, where the linear classifier stays MSE-optimal for feature activations throughout the dynamics. (III) By studying renormalized gradient flow along the central path, we derive exact dynamics that predict NC.

Weighted Training for Cross-Task Learning

- Shuxiao Chen, Koby Crammer, Hangfeng He, Dan Roth, Weijie J Su
- abstract@open-review(Oral): In this paper, we introduce Target-Aware Weighted Training (TAWT), a weighted training algorithm for cross-task learning based on minimizing a representation-based task distance between the source and target tasks. We show that TAWT is easy to implement, is computationally efficient, requires little hyperparameter tuning, and enjoys non-asymptotic learning-theoretic guarantees. The effectiveness of TAWT is corroborated through extensive experiments with BERT on four sequence tagging tasks in natural language processing (NLP), including part-of-speech (PoS) tagging, chunking, predicate detection, and named entity recognition (NER). As a byproduct, the proposed representation-based task distance allows one to reason in a theoretically principled way about several critical aspects of cross-task learning, such as the choice of the source data and the impact of fine-tuning.

iLQR-VAE : control-based learning of input-driven dynamics with applications to neural data

- Marine Schimel, Ta-Chu Kao, Kristopher T Jensen, Guillaume Hennequin
- abstract@open-review(Oral): Understanding how neural dynamics give rise to behaviour is one of the most fundamental questions in systems neuroscience. To achieve this, a common approach is to record neural populations in behaving animals, and model these data as emanating from a latent dynamical system whose state trajectories can then be related back to behavioural observations via some form of decoding. As recordings are typically performed in localized circuits that form only a part of the wider implicated network, it is important to simultaneously learn the local dynamics and infer any unobserved external input that might drive them. Here, we introduce iLQR-VAE, a novel control-based approach to variational inference in nonlinear dynamical systems, capable of learning both latent dynamics, initial conditions, and ongoing external inputs. As in recent deep learning approaches, our method is based on an input-driven sequential variational autoencoder (VAE). The main novelty lies in the use of the powerful iterative linear quadratic regulator algorithm (iLQR) in the recognition model. Optimization of the standard evidence lower-bound requires differentiating through iLQR solutions, which is made possible by recent advances in differentiable control. Importantly, having the recognition model be implicitly defined by the generative model greatly reduces the number of free parameters and allows for flexible, high-quality inference. This makes it possible for instance to evaluate the model on a single long trial after training on smaller chunks. We demonstrate the effectiveness of iLQR-VAE on a range of synthetic systems, with autonomous as well as input-driven dynamics. We further apply it to neural and behavioural recordings in non-human primates performing two different reaching tasks, and show that iLQR-VAE yields high-quality kinematic reconstructions from the neural data.

Extending the WILDS Benchmark for Unsupervised Adaptation

- Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, Sara Beery, Etienne David, Ian Stavness, Wei Guo, Jure Leskovec, Kate Saenko, Tatsunori Hashimoto, Sergey Levine, Chelsea Finn, Percy Liang
- abstract@open-review(Oral): Machine learning systems deployed in the wild are often trained on a source distribution but deployed on a different target distribution. Unlabeled data can be a powerful point of leverage for mitigating these distribution shifts, as it is frequently much more available than labeled data and can often be obtained from distributions beyond the source distribution as well. However, existing distribution shift benchmarks with unlabeled data do not reflect the breadth of scenarios that arise in real-world applications. In this work, we present the WILDS 2.0 update, which extends 8 of the 10 datasets in the WILDS benchmark of distribution shifts to include curated unlabeled data that would be realistically obtainable in deployment. These datasets span a wide range of applications (from histology to wildlife conservation), tasks (classification, regression, and detection), and modalities (photos, satellite images, microscope slides, text, molecular graphs). The update maintains consistency with the original WILDS benchmark by using identical labeled training, validation, and test sets, as well as identical evaluation metrics. We systematically benchmark state-of-the-art methods that use unlabeled data, including domain-invariant, self-training, and self-supervised methods, and show that their success on WILDS is limited. To facilitate method development, we provide an open-source package that automates data loading and contains the model architectures and methods used in this paper. Code and leaderboards are available at https://wilds.stanford.edu.

Asymmetry Learning for Counterfactually-invariant Classification in OOD Tasks

- S Chandra Mouli, Bruno Ribeiro
- abstract@open-review(Oral): Generalizing from observed to new related environments (out-of-distribution) is central to the reliability of classifiers. However, most classifiers fail to predict label Y from input X when the change in environment is due a (stochastic) input transformation $T^\text{te} \circ X'$ not observed in training, as in training we observe $T^\text{tr} \circ X'$, where X' is a hidden variable. This work argues that when the transformations in train T^tr and test T^te are (arbitrary) symmetry transformations induced by a collection of known m equivalence relations, the task of finding a robust OOD classifier can be defined as finding the simplest causal model that defines a causal connection between the target labels and the symmetry transformations that are associated with label changes. We then propose a new learning paradigm, asymmetry learning, that identifies which symmetries the classifier must break in order to correctly predict Y in both train and test. Asymmetry learning performs a causal model search that, under certain identifiability conditions, finds classifiers that perform equally well in-distribution and out-of-distribution. Finally, we show how to learn counterfactually-invariant representations with asymmetry learning in two physics tasks.

Comparing Distributions by Measuring Differences that Affect Decision Making

- Shengjia Zhao, Abhishek Sinha, Yutong He, Aidan Perreault, Jiaming Song, Stefano Ermon

- abstract@open-review(Oral): Measuring the discrepancy between two probability distributions is a fundamental problem in machine learning and statistics. We propose a new class of discrepancies based on the optimal loss for a decision task -- two distributions are different if the optimal decision loss is higher on their mixture than on each individual distribution. By suitably choosing the decision task, this generalizes the Jensen-Shannon divergence and the maximum mean discrepancy family. We apply our approach to two-sample tests, and on various benchmarks, we achieve superior test power compared to competing methods. In addition, a modeler can directly specify their preferences when comparing distributions through the decision loss. We apply this property to understanding the effects of climate change on different social and economic activities, evaluating sample quality, and selecting features targeting different decision tasks.

MIDI-DDSP: Detailed Control of Musical Performance via Hierarchical Modeling

- Yusong Wu, Ethan Manilow, Yi Deng, Rigel Swavely, Kyle Kastner, Tim Cooijmans, Aaron Courville, Cheng-Zhi Anna Huang, Jesse Engel
- abstract@open-review(Oral): Musical expression requires control of both what notes that are played, and how they are performed. Conventional audio synthesizers provide detailed expressive controls, but at the cost of realism. Black-box neural audio synthesis and concatenative samplers can produce realistic audio, but have few mechanisms for control. In this work, we introduce MIDI-DDSP a hierarchical model of musical instruments that enables both realistic neural audio synthesis and detailed user control. Starting from interpretable Differentiable Digital Signal Processing (DDSP) synthesis parameters, we infer musical notes and high-level properties of their expressive performance (such as timbre, vibrato, dynamics, and articulation). This creates a 3-level hierarchy (notes, performance, synthesis) that affords individuals the option to intervene at each level, or utilize trained priors (performance given notes, synthesis given performance) for creative assistance. Through quantitative experiments and listening tests, we demonstrate that this hierarchy can reconstruct high-fidelity audio, accurately predict performance attributes for a note sequence, independently manipulate the attributes of a given performance, and as a complete system, generate realistic audio from a novel note sequence. By utilizing an interpretable hierarchy, with multiple levels of granularity, MIDI-DDSP opens the door to assistive tools to empower individuals across a diverse range of musical experience.

Unsupervised Vision-Language Grammar Induction with Shared Structure Modeling

- Bo Wan, Wenjuan Han, Zilong Zheng, Tinne Tuytelaars
- abstract@open-review(Oral): We introduce a new task, unsupervised vision-language (VL) grammar induction. Given an image-caption pair, the goal is to extract a shared hierarchical structure for both image and language simultaneously. We argue that such structured output, grounded in both modalities, is a clear step towards the high-level understanding of multimodal information. Besides challenges existing in conventional visually grounded grammar induction tasks, VL grammar induction requires a model to capture contextual semantics and perform a fine-grained alignment. To address these challenges, we propose a novel method, CLIORA, which constructs a shared vision-language constituency tree structure with context-dependent semantics for all possible phrases in different levels of the tree. It computes a matching score between each constituent and image region, trained via contrastive learning. It integrates two levels of fusion, namely at feature-level and at score-level, so as to allow fine-grained alignment. We introduce a new evaluation metric for VL grammar induction, CCRA, and show a 3.3% improvement over a strong baseline on Flickr30k Entities. We also evaluate our model via two derived tasks, i.e., language grammar induction and phrase grounding, and improve over the state-of-the-art for both.

PiCO: Contrastive Label Disambiguation for Partial Label Learning

- Haobo Wang, Ruixuan Xiao, Yixuan Li, Lei Feng, Gang Niu, Gang Chen, Junbo Zhao
- abstract@open-review(Oral): Partial label learning (PLL) is an important problem that allows each training example to be labeled with a coarse candidate set, which well suits many real-world data annotation scenarios with label ambiguity. Despite the promise, the performance of PLL often lags behind the supervised counterpart. In this work, we bridge the gap by addressing two key research challenges in PLL---representation learning and label disambiguation---in one coherent framework. Specifically, our proposed framework PiCO consists of a contrastive learning module along with a novel class prototype-based label disambiguation algorithm. PiCO produces closely aligned representations for examples from the same classes and facilitates label disambiguation. Theoretically, we show that these two components are mutually beneficial, and can be rigorously justified from an expectation-maximization (EM) algorithm perspective. Extensive experiments demonstrate that PiCO significantly outperforms the current state-of-the-art approaches in PLL and even achieves comparable results to fully supervised learning. Code and data available: https://github.com/hbzju/PiCO.

Pyraformer: Low-Complexity Pyramidal Attention for Long-Range Time Series Modeling and Forecasting

- Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X. Liu, Schahram Dustdar
- abstract@open-review(Oral): Accurate prediction of the future given the past based on time series data is of paramount importance, since it opens the door for decision making and risk management ahead of time. In practice, the challenge is to build a flexible but parsimonious model that can capture a wide range of temporal dependencies. In this paper, we propose Pyraformer by exploring the multiresolution representation of the time series. Specifically, we introduce the pyramidal attention module (PAM) in which the inter-scale tree structure summarizes features at different resolutions and the intra-scale neighboring connections model the temporal dependencies of different ranges. Under mild conditions, the maximum length of the signal traversing path in Pyraformer is a constant (i.e., $\mathcal O(1)$) with regard to the sequence length L, while its time and space complexity scale linearly with L. Extensive numerical results show that Pyraformer typically achieves the highest prediction accuracy in both single-step and long-range forecasting tasks with the least amount of time and memory consumption, especially when the sequence is long.

Expressiveness and Approximation Properties of Graph Neural Networks

- Floris Geerts, Juan L Reutter
- abstract@open-review(Oral): Characterizing the separation power of graph neural networks (GNNs) provides an understanding of their limitations for graph learning tasks. Results regarding separation power are, however, usually geared at specific GNNs architectures, and tools for understanding arbitrary GNN architectures are generally lacking. We provide an elegant way to easily obtain bounds on the separation power of GNNs in terms of the Weisfeiler-Leman (WL) tests, which have become the yardstick to measure the separation power of GNNs. The crux is to view GNNs as expressions in a procedural tensor language describing the computations in the layers of the GNNs. Then, by a simple analysis of the obtained expressions, in terms of the number of indexes used and the nesting depth of summations, bounds on the separation power in terms of the WL-tests readily follow. We use tensor language to define Higher-Order Message-Passing Neural Networks (or k-MPNNs), a natural extension of MPNNs. Furthermore, the tensor language point of view allows for the derivation of universality results for classes of GNNs in a natural way. Our approach provides a toolbox with which GNN architecture designers can analyze the separation power of their GNNs, without needing to know the intricacies of the WL-tests. We also provide insights in what is needed to boost the separation power of GNNs.

Filtered-CoPhy: Unsupervised Learning of Counterfactual Physics in Pixel Space

- Steeven JANNY, Fabien Baradel, Natalia Neverova, Madiha Nadri, Greg Mori, Christian Wolf
- abstract@open-review(Oral): Learning causal relationships in high-dimensional data (images, videos) is a hard task, as they are often defined on low dimensional manifolds and must be extracted from complex signals dominated by appearance, lighting, textures and also spurious correlations in the data. We present a method for learning counterfactual reasoning of physical processes in pixel space, which requires the prediction of the impact of interventions on initial conditions. Going beyond the identification of structural relationships, we deal with the challenging problem of forecasting raw video over long horizons. Our method does not require the knowledge or supervision of any ground truth positions or other object or scene properties. Our model learns and acts on a suitable hybrid latent representation based on a combination of dense features, sets of 2D keypoints and an additional latent

vector per keypoint. We show that this better captures the dynamics of physical processes than purely dense or sparse representations. We introduce a new challenging and carefully designed counterfactual benchmark for predictions in pixel space and outperform strong baselines in physics-inspired ML and video prediction.

BEiT: BERT Pre-Training of Image Transformers

- Hangbo Bao, Li Dong, Songhao Piao, Furu Wei
- abstract@open-review(Oral): We introduce a self-supervised vision representation model BEiT, which stands for Bidirectional Encoder representation from Image Transformers. Following BERT developed in the natural language processing area, we propose a masked image modeling task to pretrain vision Transformers. Specifically, each image has two views in our pre-training, i.e., image patches (such as 16 x 16 pixels), and visual tokens (i.e., discrete tokens). We first ``tokenize" the original image into visual tokens. Then we randomly mask some image patches and fed them into the backbone Transformer. The pre-training objective is to recover the original visual tokens based on the corrupted image patches. After pre-training BEiT, we directly fine-tune the model parameters on downstream tasks by appending task layers upon the pretrained encoder. Experimental results on image classification and semantic segmentation show that our model achieves competitive results with previous pre-training methods.

Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution

- Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, Percy Liang
- abstract@open-review(Oral): When transferring a pretrained model to a downstream task, two popular methods are full fine-tuning (updating all the model parameters) and linear probing (updating only the last linear layer---the "head"). It is well known that fine-tuning leads to better accuracy in-distribution (ID). However, in this paper, we find that fine-tuning can achieve worse accuracy than linear probing out-of-distribution (OOD) when the pretrained features are good and the distribution shift is large. On 10 distribution shift datasets (BREEDS-Living17, BREEDS-Entity30, DomainNet, CIFAR $\to$ STL, CIFAR-10.1, FMoW, ImageNetV2, ImageNet-R, ImageNet-A, ImageNet-Sketch), fine-tuning obtains on average 2% higher accuracy ID but 7% lower accuracy OOD than linear probing. We show theoretically that this tradeoff between ID and OOD accuracy arises even in a simple setting: fine-tuning overparameterized two-layer linear networks. We prove that the OOD error of fine-tuning is high when we initialize with a fixed or random head---this is because while fine-tuning learns the head, the lower layers of the neural network change simultaneously and distort the pretrained features. Our analysis suggests that the easy two-step strategy of linear probing then full fine-tuning (LP-FT), sometimes used as a fine-tuning heuristic, combines the benefits of both fine-tuning and linear probing. Empirically, LP-FT outperforms both fine-tuning and linear probing on the above datasets (1% better ID, 10% better OOD than full fine-tuning).

StyleAlign: Analysis and Applications of Aligned StyleGAN Models

- Zongze Wu, Yotam Nitzan, Eli Shechtman, Dani Lischinski
- abstract@open-review(Oral): In this paper, we perform an in-depth study of the properties and applications of aligned generative models. We refer to two models as aligned if they share the same architecture, and one of them (the child) is obtained from the other (the parent) via fine-tuning to another domain, a common practice in transfer learning. Several works already utilize some basic properties of aligned StyleGAN models to perform image-to-image translation. Here, we perform the first detailed exploration of model alignment, also focusing on StyleGAN. First, we empirically analyze aligned models and provide answers to important questions regarding their nature. In particular, we find that the child model's latent spaces are semantically aligned with those of the parent, inheriting incredibly rich semantics, even for distant data domains such as human faces and churches. Second, equipped with this better understanding, we leverage aligned models to solve a diverse set of tasks. In addition to image translation, we demonstrate fully automatic cross-domain image morphing. We further show that zero-shot vision tasks may be performed in the child domain, while relying exclusively on supervision in the parent domain. We demonstrate qualitatively and quantitatively that our approach yields state-of-the-art results, while requiring only simple fine-tuning and inversion.

Variational Inference for Discriminative Learning with Generative Modeling of Feature Incompletion

- Kohei Miyaguchi, Takayuki Katsuki, Akira Koseki, Toshiya Iwamori
- abstract@open-review(Oral): We are concerned with the problem of distributional prediction with incomplete features: The goal is to estimate the distribution of target variables given feature vectors with some of the elements missing. A typical approach to this problem is to perform missing-value imputation and regression, simultaneously or sequentially, which we call the generative approach. Another approach is to perform regression after appropriately encoding missing values into the feature, which we call the discriminative approach. In comparison, the generative approach is more robust to the feature corruption while the discriminative approach is more favorable to maximize the performance of prediction. In this study, we propose a hybrid method to take the best of both worlds. Our method utilizes the black-box variational inference framework so that it can be applied to a wide variety of modern machine learning models, including the variational autoencoders. We also confirmed the effectiveness of the proposed method empirically.

Efficiently Modeling Long Sequences with Structured State Spaces

- Albert Gu, Karan Goel, Christopher Re
- abstract@open-review(Oral): A central goal of sequence modeling is designing a single principled model that can address sequence data across a range of modalities and tasks, particularly on long-range dependencies. Although conventional models including RNNs, CNNs, and Transformers have specialized variants for capturing long dependencies, they still struggle to scale to very long sequences of 10000 or more steps. A promising recent approach proposed modeling sequences by simulating the fundamental state space model (SSM) (x'(t) = Ax(t) + Bu(t), y(t) = Cx(t) + Du(t)), and showed that for appropriate choices of the state matrix (A), this system could handle long-range dependencies mathematically and empirically. However, this method has prohibitive computation and memory requirements, rendering it infeasible as a general sequence modeling solution. We propose the Structured State Space sequence model (S4) based on a new parameterization for the SSM, and show that it can be computed much more efficiently than prior approaches while preserving their theoretical strengths. Our technique involves conditioning (A) with a low-rank correction, allowing it to be diagonalized stably and reducing the SSM to the well-studied computation of a Cauchy kernel. S4 achieves strong empirical results across a diverse range of established benchmarks, including (i) 91\% accuracy on sequential CIFAR-10 with no data augmentation or auxiliary losses, on par with a larger 2-D ResNet, (ii) substantially closing the gap to Transformers on image and language modeling tasks, while performing generation $60\times$ faster (iii) SoTA on every task from the Long Range Arena benchmark, including solving the challenging Path-X task of length 16k that all prior work fails on, while being as efficient as all competitors.

Large Language Models Can Be Strong Differentially Private Learners

- Xuechen Li, Florian Tramer, Percy Liang, Tatsunori Hashimoto
- abstract@open-review(Oral): Differentially Private (DP) learning has seen limited success for building large deep learning models of text, and straightforward attempts at applying Differentially Private Stochastic Gradient Descent (DP-SGD) to NLP tasks have resulted in large performance drops and high computational overhead. We show that this performance drop can be mitigated with (1) the use of large pretrained language models; (2) non-standard hyperparameters that suit DP optimization; and (3) fine-tuning objectives which are aligned with the pretraining procedure. With the above, we obtain NLP models that outperform state-of-the-art DP-trained models under the same privacy budget and strong non-private baselines---by directly fine-tuning pretrained models with DP optimization on moderately-sized corpora. To address the computational challenge of running DP-SGD with large

Transformers, we propose a memory saving technique that allows clipping in DP-SGD to run without instantiating per-example gradients for any linear layer in the model. The technique enables privately training Transformers with almost the same memory cost as non-private training at a modest run-time overhead. Contrary to conventional wisdom that DP optimization fails at learning high-dimensional models (due to noise that scales with dimension) empirical results reveal that private learning with pretrained language models tends to not suffer from dimension-dependent performance degradation. Code to reproduce results can be found at https://github.com/lxuechen/private-transformers.

GeoDiff: A Geometric Diffusion Model for Molecular Conformation Generation

- Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, Jian Tang
- abstract@open-review(Oral): Predicting molecular conformations from molecular graphs is a fundamental problem in cheminformatics and drug discovery. Recently, significant progress has been achieved with machine learning approaches, especially with deep generative models. Inspired by the diffusion process in classical non-equilibrium thermodynamics where heated particles will diffuse from original states to a noise distribution, in this paper, we propose a novel generative model named GeoDiff for molecular conformation prediction. GeoDiff treats each atom as a particle and learns to directly reverse the diffusion process (i.e., transforming from a noise distribution to stable conformations) as a Markov chain. Modeling such a generation process is however very challenging as the likelihood of conformations should be roto-translational invariant. We theoretically show that Markov chains evolving with equivariant Markov kernels can induce an invariant distribution by design, and further propose building blocks for the Markov kernels to preserve the desirable equivariance property. The whole framework can be efficiently trained in an end-to-end fashion by optimizing a weighted variational lower bound to the (conditional) likelihood. Experiments on multiple benchmarks show that GeoDiff is superior or comparable to existing state-of-the-art approaches, especially on large molecules.

Frame Averaging for Invariant and Equivariant Network Design

- Omri Puny, Matan Atzmon, Edward J. Smith, Ishan Misra, Aditya Grover, Heli Ben-Hamu, Yaron Lipman
- abstract@open-review(Oral): Many machine learning tasks involve learning functions that are known to be invariant or equivariant to certain symmetries of the input data. However, it is often challenging to design neural network architectures that respect these symmetries while being expressive and computationally efficient. For example, Euclidean motion invariant/equivariant graph or point cloud neural networks. We introduce Frame Averaging (FA), a highly general purpose and systematic framework for adapting known (backbone) architectures to become invariant or equivariant to new symmetry types. Our framework builds on the well known group averaging operator that guarantees invariance or equivariance but is intractable. In contrast, we observe that for many important classes of symmetries, this operator can be replaced with an averaging operator over a small subset of the group elements, called a frame. We show that averaging over a frame guarantees exact invariance or equivariance while often being much simpler to compute than averaging over the entire group. Furthermore, we prove that FA-based models have maximal expressive power in a broad setting and in general preserve the expressive power of their backbone architectures. Using frame averaging, we propose a new class of universal Graph Neural Networks (GNNs), universal Euclidean motion invariant point cloud networks, and Euclidean motion invariant Message Passing (MP) GNNs. We demonstrate the practical effectiveness of FA on several applications including point cloud normal estimation, beyond 2-WL graph separation, and n-body dynamics prediction, achieving state-of-the-art results in all of these benchmarks.

Einops: Clear and Reliable Tensor Manipulations with Einstein-like Notation

- Alex Rogozhnikov
- abstract@open-review(Oral): Tensor computations underlie modern scientific computing and deep learning. A number of tensor frameworks emerged varying in execution model, hardware support, memory management, model definition, etc. However, tensor operations in all frameworks follow the same paradigm. Recent neural network architectures demonstrate demand for higher expressiveness of tensor operations. The current paradigm is not suited to write readable, reliable, or easy-to-modify code for multidimensional tensor manipulations. Moreover, some commonly used operations do not provide sufficient checks and can break a tensor structure. These mistakes are elusive as no tools or tests can detect them. Independently, API discrepancies complicate code transfer between frameworks. We propose einops notation: a uniform and generic way to manipulate tensor structure, that significantly improves code readability and flexibility by focusing on the structure of input and output tensors. We implement einops notation in a Python package that efficiently supports multiple widely used frameworks and provides framework-independent minimalist API for tensor manipulations.

A Fine-Grained Analysis on Distribution Shift

- Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dj Dvijotham, Ali Taylan Cemgil
- abstract@open-review(Oral): Robustness to distribution shifts is critical for deploying machine learning models in the real world. Despite this necessity, there has been little work in defining the underlying mechanisms that cause these shifts and evaluating the robustness of algorithms across multiple, different distribution shifts. To this end, we introduce a framework that enables fine-grained analysis of various distribution shifts. We provide a holistic analysis of current state-of-the-art methods by evaluating 19 distinct methods grouped into five categories across both synthetic and real-world datasets. Overall, we train more than 85K models. Our experimental framework can be easily extended to include new methods, shifts, and datasets. We find, unlike previous work (Gulrajani & Lopez-Paz, 2021), that progress has been made over a standard ERM baseline; in particular, pretraining and augmentations (learned or heuristic) offer large gains in many cases. However, the best methods are not consistent over different datasets and shifts. We will open source our experimental framework, allowing future work to evaluate new methods over multiple shifts to obtain a more complete picture of a method's effectiveness. Code is available at github.com/deepmind/distribution_shift_framework.

Open-Set Recognition: A Good Closed-Set Classifier is All You Need

- Sagar Vaze, Kai Han, Andrea Vedaldi, Andrew Zisserman
- abstract@open-review(Oral): The ability to identify whether or not a test sample belongs to one of the semantic classes in a classifier's training set is critical to practical deployment of the model. This task is termed open-set recognition (OSR) and has received significant attention in recent years. In this paper, we first demonstrate that the ability of a classifier to make the 'none-of-above' decision is highly correlated with its accuracy on the closed-set classes. We find that this relationship holds across loss objectives and architectures, and further demonstrate the trend both on the standard OSR benchmarks as well as on a large-scale ImageNet evaluation. Second, we use this correlation to boost the performance of the maximum softmax probability OSR 'baseline' by improving its closed-set accuracy, and with this strong baseline achieve state-of-the-art on a number of OSR benchmarks. Similarly, we boost the performance of the existing state-of-the-art method by improving its closed-set accuracy, but the resulting discrepancy with the strong baseline is marginal. Our third contribution is to present the 'Semantic Shift Benchmark' (SSB), which better respects the task of detecting semantic novelty, as opposed to low-level distributional shifts as tackled by neighbouring machine learning fields. On this new evaluation, we again demonstrate that there is negligible difference between the strong baseline and the existing state-of-the-art. Code available at: https://github.com/sgvaze/osr_closed_set_all_you_need.

Learning Strides in Convolutional Neural Networks

- Rachid Riad, Olivier Teboul, David Grangier, Neil Zeghidour
- abstract@open-review(Oral): Convolutional neural networks typically contain several downsampling operators, such as strided convolutions or pooling layers, that progressively reduce the resolution of intermediate representations. This provides some shift-invariance while reducing the computational complexity of the whole architecture. A critical hyperparameter of such layers is their stride: the integer factor of downsampling. As strides are not

differentiable, finding the best configuration either requires cross-validation or discrete optimization (e.g. architecture search), which rapidly become prohibitive as the search space grows exponentially with the number of downsampling layers. Hence, exploring this search space by gradient descent would allow finding better configurations at a lower computational cost. This work introduces DiffStride, the first downsampling layer with learnable strides. Our layer learns the size of a cropping mask in the Fourier domain, that effectively performs resizing in a differentiable way. Experiments on audio and image classification show the generality and effectiveness of our solution: we use DiffStride as a drop-in replacement to standard downsampling layers and outperform them. In particular, we show that introducing our layer into a ResNet-18 architecture allows keeping consistent high performance on CIFAR10, CIFAR100 and ImageNet even when training starts from poor random stride configurations. Moreover, formulating strides as learnable variables allows us to introduce a regularization term that controls the computational complexity of the architecture. We show how this regularization allows trading off accuracy for efficiency on ImageNet.

Understanding over-squashing and bottlenecks on graphs via curvature

- Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, Michael M. Bronstein
- abstract@open-review(Oral): Most graph neural networks (GNNs) use the message passing paradigm, in which node features are propagated on the input graph. Recent works pointed to the distortion of information flowing from distant nodes as a factor limiting the efficiency of message passing for tasks relying on long-distance interactions. This phenomenon, referred to as 'over-squashing', has been heuristically attributed to graph bottlenecks where the number of k-hop neighbors grows rapidly with k. We provide a precise description of the over-squashing phenomenon in GNNs and analyze how it arises from bottlenecks in the graph. For this purpose, we introduce a new edge-based combinatorial curvature and prove that negatively curved edges are responsible for the over-squashing issue. We also propose and experimentally test a curvature-based graph rewiring method to alleviate the over-squashing.

Diffusion-Based Voice Conversion with Fast Maximum Likelihood Sampling Scheme

- Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, Mikhail Sergeevich Kudinov, Jiansheng Wei
- abstract@open-review(Oral): Voice conversion is a common speech synthesis task which can be solved in different ways depending on a particular real-world scenario. The most challenging one often referred to as one-shot many-to-many voice conversion consists in copying target voice from only one reference utterance in the most general case when both source and target speakers do not belong to the training dataset. We present a scalable high-quality solution based on diffusion probabilistic modeling and demonstrate its superior quality compared to state-of-the-art one-shot voice conversion approaches. Moreover, focusing on real-time applications, we investigate general principles which can make diffusion models faster while keeping synthesis quality at a high level. As a result, we develop a novel Stochastic Differential Equations solver suitable for various diffusion model types and generative tasks as shown through empirical studies and justify it by theoretical analysis.

Resolving Training Biases via Influence-based Data Relabeling

- Shuming Kong, Yanyan Shen, Linpeng Huang
- abstract@open-review(Oral): The performance of supervised learning methods easily suffers from the training bias issue caused by train-test distribution mismatch or label noise. Influence function is a technique that estimates the impacts of a training sample on the modelâ€™s predictions. Recent studies on \emph{data resampling} have employed influence functions to identify \emph{harmful} training samples that will degrade model's test performance. They have shown that discarding or downweighting the identified harmful training samples is an effective way to resolve training biases. In this work, we move one step forward and propose an influence-based relabeling framework named RDIA for reusing harmful training samples toward better model performance. To achieve this, we use influence functions to estimate how relabeling a training sample would affect model's test performance and further develop a novel relabeling function R. We theoretically prove that applying R to relabel harmful training samples allows the model to achieve lower test loss than simply discarding them for any classification tasks using cross-entropy loss. Extensive experiments on ten real-world datasets demonstrate RDIA outperforms the state-of-the-art data resampling methods and improves model's robustness against label noise.

Representational Continuity for Unsupervised Continual Learning

- Divyam Madaan, Jaehong Yoon, Yuanchun Li, Yunxin Liu, Sung Ju Hwang
- abstract@open-review(Oral): Continual learning (CL) aims to learn a sequence of tasks without forgetting the previously acquired knowledge. However, recent CL advances are restricted to supervised continual learning (SCL) scenarios. Consequently, they are not scalable to real-world applications where the data distribution is often biased and unannotated. In this work, we focus on unsupervised continual learning (UCL), where we learn the feature representations on an unlabelled sequence of tasks and show that reliance on annotated data is not necessary for continual learning. We conduct a systematic study analyzing the learned feature representations and show that unsupervised visual representations are surprisingly more robust to catastrophic forgetting, consistently achieve better performance, and generalize better to out-of-distribution tasks than SCL. Furthermore, we find that UCL achieves a smoother loss landscape through qualitative analysis of the learned representations and learns meaningful feature representations. Additionally, we propose Lifelong Unsupervised Mixup (LUMP), a simple yet effective technique that interpolates between the current task and previous tasks' instances to alleviate catastrophic forgetting for unsupervised representations.

Vision-Based Manipulators Need to Also See from Their Hands

- Kyle Hsu, Moo Jin Kim, Rafael Rafailov, Jiajun Wu, Chelsea Finn
- abstract@open-review(Oral): We study how the choice of visual perspective affects learning and generalization in the context of physical manipulation from raw sensor observations. Compared with the more commonly used global third-person perspective, a hand-centric (eye-in-hand) perspective affords reduced observability, but we find that it consistently improves training efficiency and out-of-distribution generalization. These benefits hold across a variety of learning algorithms, experimental settings, and distribution shifts, and for both simulated and real robot apparatuses. However, this is only the case when hand-centric observability is sufficient; otherwise, including a third-person perspective is necessary for learning, but also harms out-of-distribution generalization. To mitigate this, we propose to regularize the third-person information stream via a variational information bottleneck. On six representative manipulation tasks with varying hand-centric observability adapted from the Meta-World benchmark, this results in a state-of-the-art reinforcement learning agent operating from both perspectives improving its out-of-distribution generalization on every task. While some practitioners have long put cameras in the hands of robots, our work systematically analyzes the benefits of doing so and provides simple and broadly applicable insights for improving end-to-end learned vision-based robotic manipulation.

Meta-Learning with Fewer Tasks through Task Interpolation

- Huaxiu Yao, Linjun Zhang, Chelsea Finn
- abstract@open-review(Oral): Meta-learning enables algorithms to quickly learn a newly encountered task with just a few labeled examples by transferring previously learned knowledge. However, the bottleneck of current meta-learning algorithms is the requirement of a large number of meta-training tasks, which may not be accessible in real-world scenarios. To address the challenge that available tasks may not densely sample the space of tasks, we propose to augment the task set through interpolation. By meta-learning with task interpolation (MLTI), our approach effectively generates additional tasks by randomly sampling a pair of tasks and interpolating the corresponding features and labels. Under both gradient-based and metric-based meta-learning settings, our theoretical analysis shows MLTI corresponds to a data-adaptive meta-regularization and further improves the generalization. Empirically, in our experiments on eight datasets from diverse domains including image recognition, pose prediction, molecule property prediction, and medical image

classification, we find that the proposed general MLTI framework is compatible with representative meta-learning algorithms and consistently outperforms other state-of-the-art strategies.

DISCOVERING AND EXPLAINING THE REPRESENTATION BOTTLENECK OF DNNS

- Huiqi Deng, Qihan Ren, Hao Zhang, Quanshi Zhang
- abstract@open-review(Oral): This paper explores the bottleneck of feature representations of deep neural networks (DNNs), from the perspective of the complexity of interactions between input variables encoded in DNNs. To this end, we focus on the multi-order interaction between input variables, where the order represents the complexity of interactions. We discover that a DNN is more likely to encode both too simple and too complex interactions, but usually fails to learn interactions of intermediate complexity. Such a phenomenon is widely shared by different DNNs for different tasks. This phenomenon indicates a cognition gap between DNNs and humans, and we call it a representation bottleneck. We theoretically prove the underlying reason for the representation bottleneck. Furthermore, we propose losses to encourage/penalize the learning of interactions of specific complexities, and analyze the representation capacities of interactions of different complexities. The code is available at https://github.com/Nebularaid2000/bottleneck.

Sparse Communication via Mixed Distributions

- AntÃ³nio Farinhas, Wilker Aziz, Vlad Niculae, Andre Martins
- abstract@open-review(Oral): Neural networks and other machine learning models compute continuous representations, while humans communicate mostly through discrete symbols. Reconciling these two forms of communication is desirable for generating human-readable interpretations or learning discrete latent variable models, while maintaining end-to-end differentiability. Some existing approaches (such as the Gumbel-Softmax transformation) build continuous relaxations that are discrete approximations in the zero-temperature limit, while others (such as sparsemax transformations and the Hard Concrete distribution) produce discrete/continuous hybrids. In this paper, we build rigorous theoretical foundations for these hybrids, which we call "mixed random variables." Our starting point is a new "direct sum" base measure defined on the face lattice of the probability simplex. From this measure, we introduce new entropy and Kullback-Leibler divergence functions that subsume the discrete and differential cases and have interpretations in terms of code optimality. Our framework suggests two strategies for representing and sampling mixed random variables, an extrinsic ("sample-and-project'â€™) and an intrinsic one (based on face stratification). We experiment with both approaches on an emergent communication benchmark and on modeling MNIST and Fashion-MNIST data with variational auto-encoders with mixed latent variables.

Finetuned Language Models are Zero-Shot Learners

- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, Quoc V Le
- abstract@open-review(Oral): This paper explores a simple method for improving the zero-shot learning abilities of language models. We show that instruction tuningâ€”finetuning language models on a collection of datasets described via instructionsâ€”substantially improves zero-shot performance on unseen tasks. We take a 137B parameter pretrained language model and instruction tune it on over 60 NLP datasets verbalized via natural language instruction templates. We evaluate this instruction-tuned model, which we call FLAN, on unseen task types. FLAN substantially improves the performance of its unmodified counterpart and surpasses zero-shot 175B GPT-3 on 20 of 25 datasets that we evaluate. FLAN even outperforms few-shot GPT-3 by a large margin on ANLI, RTE, BoolQ, AI2-ARC, OpenbookQA, and StoryCloze. Ablation studies reveal that number of finetuning datasets, model scale, and natural language instructions are key to the success of instruction tuning.

F8Net: Fixed-Point 8-bit Only Multiplication for Network Quantization

- Qing Jin, Jian Ren, Richard Zhuang, Sumant Hanumante, Zhengang Li, Zhiyu Chen, Yanzhi Wang, Kaiyuan Yang, Sergey Tulyakov
- abstract@open-review(Oral): Neural network quantization is a promising compression technique to reduce memory footprint and save energy consumption, potentially leading to real-time inference. However, there is a performance gap between quantized and full-precision models. To reduce it, existing quantization approaches require high-precision INT32 or full-precision multiplication during inference for scaling or dequantization. This introduces a noticeable cost in terms of memory, speed, and required energy. To tackle these issues, we present F8Net, a novel quantization framework consisting in only ï¬ xed-point 8-bit multiplication. To derive our method, we ï¬ rst discuss the advantages of ï¬ xed-point multiplication with different formats of ï¬ xed-point numbers and study the statistical behavior of the associated ï¬ xed-point numbers. Second, based on the statistical and algorithmic analysis, we apply different ï¬ xed-point formats for weights and activations of different layers. We introduce a novel algorithm to automatically determine the right format for each layer during training. Third, we analyze a previous quantization algorithmâ€”parameterized clipping activation (PACT)â€”and reformulate it using ï¬ xed-point arithmetic. Finally, we unify the recently proposed method for quantization ï¬ ne-tuning and our ï¬ xed-point approach to show the potential of our method. We verify F8Net on ImageNet for MobileNet V1/V2 and ResNet18/50. Our approach achieves comparable and better performance, when compared not only to existing quantization techniques with INT32 multiplication or ï¬,oating point arithmetic, but also to the full-precision counterparts, achieving state-of-the-art performance.

Transform2Act: Learning a Transform-and-Control Policy for Efficient Agent Design

- Ye Yuan, Yuda Song, Zhengyi Luo, Wen Sun, Kris M. Kitani
- abstract@open-review(Oral): An agent's functionality is largely determined by its design, i.e., skeletal structure and joint attributes (e.g., length, size, strength). However, finding the optimal agent design for a given function is extremely challenging since the problem is inherently combinatorial and the design space is prohibitively large. Additionally, it can be costly to evaluate each candidate design which requires solving for its optimal controller. To tackle these problems, our key idea is to incorporate the design procedure of an agent into its decision-making process. Specifically, we learn a conditional policy that, in an episode, first applies a sequence of transform actions to modify an agent's skeletal structure and joint attributes, and then applies control actions under the new design. To handle a variable number of joints across designs, we use a graph-based policy where each graph node represents a joint and uses message passing with its neighbors to output joint-specific actions. Using policy gradient methods, our approach enables joint optimization of agent design and control as well as experience sharing across different designs, which improves sample efficiency substantially. Experiments show that our approach, Transform2Act, outperforms prior methods significantly in terms of convergence speed and final performance. Notably, Transform2Act can automatically discover plausible designs similar to giraffes, squids, and spiders. Code and videos are available at https://sites.google.com/view/transform2act.

ProtoRes: Proto-Residual Network for Pose Authoring via Learned Inverse Kinematics

- Boris N. Oreshkin, Florent Bocquelet, Felix G. Harvey, Bay Raitt, Dominic Laflamme
- abstract@open-review(Oral): Our work focuses on the development of a learnable neural representation of human pose for advanced AI assisted animation tooling. Specifically, we tackle the problem of constructing a full static human pose based on sparse and variable user inputs (e.g. locations and/or orientations of a subset of body joints). To solve this problem, we propose a novel neural architecture that combines residual connections with prototype encoding of a partially specified pose to create a new complete pose from the learned latent space. We show that our architecture outperforms a baseline based on Transformer, both in terms of accuracy and computational efficiency. Additionally, we develop a user interface to integrate our neural model in Unity, a real-time 3D development platform. Furthermore, we introduce two new datasets representing the static human pose modeling problem, based on high-quality human motion capture data, which will be released publicly along with model code.

Hyperparameter Tuning with Renyi Differential Privacy

- Nicolas Papernot, Thomas Steinke
- abstract@open-review(Oral): For many differentially private algorithms, such as the prominent noisy stochastic gradient descent (DP-SGD), the analysis needed to bound the privacy leakage of a single training run is well understood. However, few studies have reasoned about the privacy leakage resulting from the multiple training runs needed to fine tune the value of the training algorithmâ€™s hyperparameters. In this work, we first illustrate how simply setting hyperparameters based on non-private training runs can leak private information. Motivated by this observation, we then provide privacy guarantees for hyperparameter search procedures within the framework of Renyi Differential Privacy. Our results improve and extend the work of Liu and Talwar (STOC 2019). Our analysis supports our previous observation that tuning hyperparameters does indeed leak private information, but we prove that, under certain assumptions, this leakage is modest, as long as each candidate training run needed to select hyperparameters is itself differentially private.

Real-Time Neural Voice Camouflage

- Mia Chiquier, Chengzhi Mao, Carl Vondrick
- abstract@open-review(Oral): Automatic speech recognition systems have created exciting possibilities for applications, however they also enable opportunities for systematic eavesdropping.We propose a method to camouflage a person's voice from these systems without inconveniencing the conversation between people in the room. Standard adversarial attacks are not effective in real-time streaming situations because the characteristics of the signal will have changed by the time the attack is executed. We introduce predictive adversarial attacks, which achieves real-time performance by forecasting the attack vector that will be the most effective in the future. Under real-time constraints, our method jams the established speech recognition system DeepSpeech 3.9x more than online projected gradient descent as measured through word error rate, and 6.6x more as measured through character error rate. We furthermore demonstrate our approach is practically effective in realistic environments with complex scene geometries.

CycleMLP: A MLP-like Architecture for Dense Prediction

- Shoufa Chen, Enze Xie, Chongjian GE, Runjian Chen, Ding Liang, Ping Luo
- abstract@open-review(Oral): This paper presents a simple MLP-like architecture, CycleMLP, which is a versatile backbone for visual recognition and dense predictions. As compared to modern MLP architectures, e.g. , MLP-Mixer, ResMLP, and gMLP, whose architectures are correlated to image size and thus are infeasible in object detection and segmentation, CycleMLP has two advantages compared to modern approaches. (1) It can cope with various image sizes. (2) It achieves linear computational complexity to image size by using local windows. In contrast, previous MLPs have $O(N^2)$ computations due to fully spatial connections. We build a family of models which surpass existing MLPs and even state-of-the-art Transformer-based models, e.g. Swin Transformer, while using fewer parameters and FLOPs. We expand the MLP-like modelsâ€™ applicability, making them a versatile backbone for dense prediction tasks. CycleMLP achieves competitive results on object detection, instance segmentation, and semantic segmentation. In particular, CycleMLP-Tiny outperforms Swin-Tiny by 1.3% mIoU on ADE20K dataset with fewer FLOPs. Moreover, CycleMLP also shows excellent zero-shot robustness on ImageNet-C dataset.

Analytic-DPM: an Analytic Estimate of the Optimal Reverse Variance in Diffusion Probabilistic Models

- Fan Bao, Chongxuan Li, Jun Zhu, Bo Zhang
- abstract@open-review(Oral): Diffusion probabilistic models (DPMs) represent a class of powerful generative models. Despite their success, the inference of DPMs is expensive since it generally needs to iterate over thousands of timesteps. A key problem in the inference is to estimate the variance in each timestep of the reverse process. In this work, we present a surprising result that both the optimal reverse variance and the corresponding optimal KL divergence of a DPM have analytic forms w.r.t. its score function. Building upon it, we propose \textit{Analytic-DPM}, a training-free inference framework that estimates the analytic forms of the variance and KL divergence using the Monte Carlo method and a pretrained score-based model. Further, to correct the potential bias caused by the score-based model, we derive both lower and upper bounds of the optimal variance and clip the estimate for a better result. Empirically, our analytic-DPM improves the log-likelihood of various DPMs, produces high-quality samples, and meanwhile enjoys a $20\times$ to $80\times$ speed up.

RISP: Rendering-Invariant State Predictor with Differentiable Simulation and Rendering for Cross-Domain Parameter Estimation

- Pingchuan Ma, Tao Du, Joshua B. Tenenbaum, Wojciech Matusik, Chuang Gan
- abstract@open-review(Oral): This work considers identifying parameters characterizing a physical system's dynamic motion directly from a video whose rendering configurations are inaccessible. Existing solutions require massive training data or lack generalizability to unknown rendering configurations. We propose a novel approach that marries domain randomization and differentiable rendering gradients to address this problem. Our core idea is to train a rendering-invariant state-prediction (RISP) network that transforms image differences into state differences independent of rendering configurations, e.g., lighting, shadows, or material reflectance. To train this predictor, we formulate a new loss on rendering variances using gradients from differentiable rendering. Moreover, we present an efficient, second-order method to compute the gradients of this loss, allowing it to be integrated seamlessly into modern deep learning frameworks. We evaluate our method in rigid-body and deformable-body simulation environments using four tasks: state estimation, system identification, imitation learning, and visuomotor control. We further demonstrate the efficacy of our approach on a real-world example: inferring the state and action sequences of a quadrotor from a video of its motion sequences. Compared with existing methods, our approach achieves significantly lower reconstruction errors and has better generalizability among unknown rendering configurations.

The Information Geometry of Unsupervised Reinforcement Learning

- Benjamin Eysenbach, Ruslan Salakhutdinov, Sergey Levine
- abstract@open-review(Oral): How can a reinforcement learning (RL) agent prepare to solve downstream tasks if those tasks are not known a priori? One approach is unsupervised skill discovery, a class of algorithms that learn a set of policies without access to a reward function. Such algorithms bear a close resemblance to representation learning algorithms (e.g., contrastive learning) in supervised learning, in that both are pretraining algorithms that maximize some approximation to a mutual information objective. While prior work has shown that the set of skills learned by such methods can accelerate downstream RL tasks, prior work offers little analysis into whether these skill learning algorithms are optimal, or even what notion of optimality would be appropriate to apply to them. In this work, we show that unsupervised skill discovery algorithms based on mutual information maximization do not learn skills that are optimal for every possible reward function. However, we show that the distribution over skills provides an optimal initialization minimizing regret against adversarially-chosen reward functions, assuming a certain type of adaptation procedure. Our analysis also provides a geometric perspective on these skill learning methods.

Language modeling via stochastic processes

- Rose E Wang, Esin Durmus, Noah Goodman, Tatsunori Hashimoto
- abstract@open-review(Oral): Modern language models can generate high-quality short texts. However, they often meander or are incoherent when generating longer texts. These issues arise from the next-token-only language modeling objective. To address these issues, we introduce Time Control (TC), a language model that implicitly plans via a latent stochastic process. TC does this by learning a representation which maps the dynamics of how

text changes in a document to the dynamics of a stochastic process of interest. Using this representation, the language model can generate text by first implicitly generating a document plan via a stochastic process, and then generating text that is consistent with this latent plan. Compared to domain-specific methods and fine-tuning GPT2 across a variety of text domains, TC improves performance on text infilling and discourse coherence. On long text generation settings, TC preserves the text structure both in terms of ordering (up to +40% better) and text length consistency (up to +17% better). Human evaluators also prefer TC's output 28.6% more than the baselines.

Wiring Up Vision: Minimizing Supervised Synaptic Updates Needed to Produce a Primate Ventral Stream

- Franziska Geiger, Martin Schrimpf, Tiago Marques, James J. DiCarlo
- abstract@open-review(Spotlight): After training on large datasets, certain deep neural networks are surprisingly good models of the neural mechanisms of adult primate visual object recognition. Nevertheless, these models are considered poor models of the development of the visual system because they posit millions of sequential, precisely coordinated synaptic updates, each based on a labeled image. While ongoing research is pursuing the use of unsupervised proxies for labels, we here explore a complementary strategy of reducing the required number of supervised synaptic updates to produce an adult-like ventral visual stream (as judged by the match to V1, V2, V4, IT, and behavior). Such models might require less precise machinery and energy expenditure to coordinate these updates and would thus move us closer to viable neuroscientific hypotheses about how the visual system wires itself up. Relative to standard model training on labeled images in ImageNet, we here demonstrate that the total number of supervised weight updates can be substantially reduced using three complementary strategies: First, we find that only 2% of supervised updates (epochs and images) are needed to achieve 80% of the match to adult ventral stream. Specifically, training benefits predictions of higher visual cortex the most whereas early visual cortex predictions only improve marginally over the course of training. Second, by improving the random distribution of synaptic connectivity, we find that 54% of the brain match can already be achieved â€œat birth" (i.e. no training at all). Third, we find that, by training only 5% of model synapses, we can still achieve nearly 80% of the match to the ventral stream. This approach further improves on ImageNet performance over previous attempts in computer vision of minimizing trained components without substantially increasing the relative number of trained parameters. These results reflect first steps in modeling not just primate adult visual processing during inference, but also how the ventral visual stream might be "wired up" by evolution (a model's "birth" state) and by developmental learning (a model's updates based on visual experience).

Dynamics-Aware Comparison of Learned Reward Functions

- Blake Wulfe, Logan Michael Ellis, Jean Mercat, Rowan Thomas McAllister, Adrien Gaidon
- abstract@open-review(Spotlight): The ability to learn reward functions plays an important role in enabling the deployment of intelligent agents in the real world. However, $\textit{comparing}$ reward functions, for example as a means of evaluating reward learning methods, presents a challenge. Reward functions are typically compared by considering the behavior of optimized policies, but this approach conflates deficiencies in the reward function with those of the policy search algorithm used to optimize it. To address this challenge, Gleave et al. (2020) propose the Equivalent-Policy Invariant Comparison (EPIC) distance. EPIC avoids policy optimization, but in doing so requires computing reward values at transitions that may be impossible under the system dynamics. This is problematic for learned reward functions because it entails evaluating them outside of their training distribution, resulting in inaccurate reward values that we show can render EPIC ineffective at comparing rewards. To address this problem, we propose the Dynamics-Aware Reward Distance (DARD), a new reward pseudometric. DARD uses an approximate transition model of the environment to transform reward functions into a form that allows for comparisons that are invariant to reward shaping while only evaluating reward functions on transitions close to their training distribution. Experiments in simulated physical domains demonstrate that DARD enables reliable reward comparisons without policy optimization and is significantly more predictive than baseline methods of downstream policy performance when dealing with learned reward functions.

Learning Hierarchical Structures with Differentiable Nondeterministic Stacks

- Brian DuSell, David Chiang
- abstract@open-review(Spotlight): Learning hierarchical structures in sequential data -- from simple algorithmic patterns to natural language -- in a reliable, generalizable way remains a challenging problem for neural language models. Past work has shown that recurrent neural networks (RNNs) struggle to generalize on held-out algorithmic or syntactic patterns without supervision or some inductive bias. To remedy this, many papers have explored augmenting RNNs with various differentiable stacks, by analogy with finite automata and pushdown automata (PDAs). In this paper, we improve the performance of our recently proposed Nondeterministic Stack RNN (NS-RNN), which uses a differentiable data structure that simulates a nondeterministic PDA, with two important changes. First, the model now assigns unnormalized positive weights instead of probabilities to stack actions, and we provide an analysis of why this improves training. Second, the model can directly observe the state of the underlying PDA. Our model achieves lower cross-entropy than all previous stack RNNs on five context-free language modeling tasks (within 0.05 nats of the information-theoretic lower bound), including a task on which the NS-RNN previously failed to outperform a deterministic stack RNN baseline. Finally, we propose a restricted version of the NS-RNN that incrementally processes infinitely long sequences, and we present language modeling results on the Penn Treebank.

Sampling with Mirrored Stein Operators

- Jiaxin Shi, Chang Liu, Lester Mackey
- abstract@open-review(Spotlight): We introduce a new family of particle evolution samplers suitable for constrained domains and non-Euclidean geometries. Stein Variational Mirror Descent and Mirrored Stein Variational Gradient Descent minimize the Kullback-Leibler (KL) divergence to constrained target distributions by evolving particles in a dual space defined by a mirror map. Stein Variational Natural Gradient exploits non-Euclidean geometry to more efficiently minimize the KL divergence to unconstrained targets. We derive these samplers from a new class of mirrored Stein operators and adaptive kernels developed in this work. We demonstrate that these new samplers yield accurate approximations to distributions on the simplex, deliver valid confidence intervals in post-selection inference, and converge more rapidly than prior methods in large-scale unconstrained posterior inference. Finally, we establish the convergence of our new procedures under verifiable conditions on the target distribution.

Planning in Stochastic Environments with a Learned Model

- Ioannis Antonoglou, Julian Schrittwieser, Sherjil Ozair, Thomas K Hubert, David Silver
- abstract@open-review(Spotlight): Model-based reinforcement learning has proven highly successful. However, learning a model in isolation from its use during planning is problematic in complex environments. To date, the most effective techniques have instead combined value-equivalent model learning with powerful tree-search methods. This approach is exemplified by MuZero, which has achieved state-of-the-art performance in a wide range of domains, from board games to visually rich environments, with discrete and continuous action spaces, in online and offline settings. However, previous instantiations of this approach were limited to the use of deterministic models. This limits their performance in environments that are inherently stochastic, partially observed, or so large and complex that they appear stochastic to a finite agent. In this paper we extend this approach to learn and plan with stochastic models. Specifically, we introduce a new algorithm, Stochastic MuZero, that learns a stochastic model incorporating afterstates, and uses this model to perform a stochastic tree search. Stochastic MuZero matched or exceeded the state of the art in a set of canonical single and multi-agent environments, including 2048 and backgammon, while maintaining the same performance as standard MuZero in the game of Go.

RotoGrad: Gradient Homogenization in Multitask Learning

- AdriÃ¡n Javaloy, Isabel Valera

- abstract@open-review(Spotlight): Multitask learning is being increasingly adopted in applications domains like computer vision and reinforcement learning. However, optimally exploiting its advantages remains a major challenge due to the effect of negative transfer. Previous works have tracked down this issue to the disparities in gradient magnitudes and directions across tasks, when optimizing the shared network parameters. While recent work has acknowledged that negative transfer is a two-fold problem, existing approaches fall short as they only focus on either homogenizing the gradient magnitude across tasks; or greedily change the gradient directions, overlooking future conflicts. In this work, we introduce RotoGrad, an algorithm that tackles negative transfer as a whole: it jointly homogenizes gradient magnitudes and directions, while ensuring training convergence. We show that RotoGrad outperforms competing methods in complex problems, including multi-label classification in CelebA and computer vision tasks in the NYUv2 dataset. A Pytorch implementation can be found in https://github.com/adrianjav/rotograd.

On Improving Adversarial Transferability of Vision Transformers

- Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Fahad Khan, Fatih Porikli
- abstract@open-review(Spotlight): Vision transformers (ViTs) process input images as sequences of patches via self-attention; a radically different architecture than convolutional neural networks (CNNs). This makes it interesting to study the adversarial feature space of ViT models and their transferability. In particular, we observe that adversarial patterns found via conventional adversarial attacks show very \emph{low} black-box transferability even for large ViT models. We show that this phenomenon is only due to the sub-optimal attack procedures that do not leverage the true representation potential of ViTs. A deep ViT is composed of multiple blocks, with a consistent architecture comprising of self-attention and feed-forward layers, where each block is capable of independently producing a class token. Formulating an attack using only the last class token (conventional approach) does not directly leverage the discriminative information stored in the earlier tokens, leading to poor adversarial transferability of ViTs.Using the compositional nature of ViT models, we enhance transferability of existing attacks by introducing two novel strategies specific to the architecture of ViT models. \emph{(i) Self-Ensemble:} We propose a method to find multiple discriminative pathways by dissecting a single ViT model into an ensemble of networks. This allows explicitly utilizing class-specific information at each ViT block. \emph{(ii) Token Refinement:} We then propose to refine the tokens to further enhance the discriminative capacity at each block of ViT.Our token refinement systematically combines the class tokens with structural information preserved within the patch tokens. An adversarial attack when applied to such refined tokens within the ensemble of classifiers found in a single vision transformer has significantly higher transferability and thereby brings out the true generalization potential of the ViT's adversarial space. Code: https://t.ly/hBbW.

On Predicting Generalization using GANs

- Yi Zhang, Arushi Gupta, Nikunj Saunshi, Sanjeev Arora
- abstract@open-review(Spotlight): Research on generalization bounds for deep networks seeks to give ways to predict test error using just the training dataset and the network parameters. While generalization bounds can give many insights about architecture design, training algorithms etc., what they do not currently do is yield good predictions for actual test error. A recently introduced Predicting Generalization in Deep Learning competition aims to encourage discovery of methods to better predict test error. The current paper investigates a simple idea: can test error be predicted using {\em synthetic data,} produced using a Generative Adversarial Network (GAN) that was trained on the same training dataset? Upon investigating several GAN models and architectures, we find that this turns out to be the case.

In fact, using GANs pre-trained on standard datasets, the test error can be predicted without requiring any additional hyper-parameter tuning. This result is surprising because GANs have well-known limitations (e.g. mode collapse) and are known to not learn the data distribution accurately. Yet the generated samples are good enough to substitute for test data. Several additional experiments are presented to explore reasons why GANs do well at this task. In addition to a new approach for predicting generalization, the counter-intuitive phenomena presented in our work may also call for a better understanding of GANs' strengths and limitations.

On the Connection between Local Attention and Dynamic Depth-wise Convolution

- Qi Han, Zejia Fan, Qi Dai, Lei Sun, Ming-Ming Cheng, Jiaying Liu, Jingdong Wang
- abstract@open-review(Spotlight): Vision Transformer (ViT) attains state-of-the-art performance in visual recognition, and the variant, Local Vision Transformer, makes further improvements. The major component in Local Vision Transformer, local attention, performs the attention separately over small local windows. We rephrase local attention as a channel-wise locally-connected layer and analyze it from two network regularization manners, sparse connectivity and weight sharing, as well as dynamic weight computation. We point out that local attention resembles depth-wise convolution and its dynamic variants in sparse connectivity: there is no connection across channels, and each position is connected to the positions within a small local window. The main differences lie in (i) weight sharing - depth-wise convolution shares connection weights (kernel weights) across spatial positions and attention shares the connection weights across channels, and (ii) dynamic weight computation manners - local attention is based on dot-products between pairwise positions in the local window, and dynamic convolution is based on linear projections conducted on the center representation or the globally pooled representation. The connection between local attention and dynamic depth-wise convolution is empirically verified by the ablation study about weight sharing and dynamic weight computation in Local Vision Transformer and (dynamic) depth-wise convolution. We empirically observe that the models based on depth-wise convolution and the dynamic variants with lower computation complexity perform on-par with or slightly better than Swin Transformer, an instance of Local Vision Transformer, for ImageNet classification, COCO object detection and ADE semantic segmentation. Code is available at https://github.com/Atten4Vis/DemystifyLocalViT.

Strength of Minibatch Noise in SGD

- Liu Ziyin, Kangqiao Liu, Takashi Mori, Masahito Ueda
- abstract@open-review(Spotlight): The noise in stochastic gradient descent (SGD), caused by minibatch sampling, is poorly understood despite its practical importance in deep learning. This work presents the first systematic study of the SGD noise and fluctuations close to a local minimum. We first analyze the SGD noise in linear regression in detail and then derive a general formula for approximating SGD noise in different types of minima. For application, our results (1) provide insight into the stability of training a neural network, (2) suggest that a large learning rate can help generalization by introducing an implicit regularization, (3) explain why the linear learning rate-batchsize scaling law fails at a large learning rate or at a small batchsize and (4) can provide an understanding of how discrete-time nature of SGD affects the recently discovered power-law phenomenon of SGD.

Learning more skills through optimistic exploration

- DJ Strouse, Kate Baumli, David Warde-Farley, Volodymyr Mnih, Steven Stenberg Hansen
- abstract@open-review(Spotlight): Unsupervised skill learning objectives (Eysenbach et al., 2019; Gregor et al., 2016) allow agents to learn rich repertoires of behavior in the absence of extrinsic rewards. They work by simultaneously training a policy to produce distinguishable latent-conditioned trajectories, and a discriminator to evaluate distinguishability by trying to infer latents from trajectories. The hope is for the agent to explore and master the environment by encouraging each skill (latent) to reliably reach different states. However, an inherent exploration problem lingers: when a novel state is actually encountered, the discriminator will necessarily not have seen enough training data to produce accurate and confident skill classifications, leading to low intrinsic reward for the agent and effective penalization of the sort of exploration needed to actually maximize the objective. To combat this inherent pessimism towards exploration, we derive an information gain auxiliary objective that involves training an ensemble of discriminators and rewarding the policy for their disagreement. Our objective directly estimates the epistemic uncertainty that comes from the discriminator not having seen enough training examples, thus providing an intrinsic reward more tailored to the true objective compared to pseudocount-based methods (Burda et al.,

2019). We call this exploration bonus discriminator disagreement intrinsic reward, or DISDAIN. We demonstrate empirically that DISDAIN improves skill learning both in a tabular grid world (Four Rooms) and the 57 games of the Atari Suite (from pixels). Thus, we encourage researchers to treat pessimism with DISDAIN.

Reinforcement Learning under a Multi-agent Predictive State Representation Model: Method and Theory

- Zhi Zhang, Zhuoran Yang, Han Liu, Pratap Tokekar, Furong Huang
- abstract@open-review(Spotlight): We study reinforcement learning for partially observable multi-agent systems where each agent only has access to its own observation and reward and aims to maximize its cumulative rewards. To handle partial observations, we propose graph-assisted predictive state representations (GAPSR), a scalable multi-agent representation learning framework that leverages the agent connectivity graphs to aggregate local representations computed by each agent. In addition, our representations are readily able to incorporate dynamic interaction graphs and kernel space embeddings of the predictive states, and thus have strong flexibility and representation power. Based on GAPSR, we propose an end-to-end MARL algorithm that simultaneously infers the predictive representations and uses the representations as the input of a policy optimization algorithm. Empirically, we demonstrate the efficacy of the proposed algorithm provided on both a MAMuJoCo robotic learning experiment and a multi-agent particle learning environment.

Adversarial Support Alignment

- Shangyuan Tong, Timur Garipov, Yang Zhang, Shiyu Chang, Tommi S. Jaakkola
- abstract@open-review(Spotlight): We study the problem of aligning the supports of distributions. Compared to the existing work on distribution alignment, support alignment does not require the densities to be matched. We propose symmetric support difference as a divergence measure to quantify the mismatch between supports. We show that select discriminators (e.g. discriminator trained for Jensen-Shannon divergence) are able to map support differences as support differences in their one-dimensional output space. Following this result, our method aligns supports by minimizing a symmetrized relaxed optimal transport cost in the discriminator 1D space via an adversarial process. Furthermore, we show that our approach can be viewed as a limit of existing notions of alignment by increasing transportation assignment tolerance. We quantitatively evaluate the method across domain adaptation tasks with shifts in label distributions. Our experiments show that the proposed method is more robust against these shifts than other alignment-based baselines.

GreaseLM: Graph REASoning Enhanced Language Models

- Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, Jure Leskovec
- abstract@open-review(Spotlight): Answering complex questions about textual narratives requires reasoning over both stated context and the world knowledge that underlies it. However, pretrained language models (LM), the foundation of most modern QA systems, do not robustly represent latent relationships between concepts, which is necessary for reasoning. While knowledge graphs (KG) are often used to augment LMs with structured representations of world knowledge, it remains an open question how to effectively fuse and reason over the KG representations and the language context, which provides situational constraints and nuances. In this work, we propose GreaseLM, a new model that fuses encoded representations from pretrained LMs and graph neural networks over multiple layers of modality interaction operations. Information from both modalities propagates to the other, allowing language context representations to be grounded by structured world knowledge, and allowing linguistic nuances (e.g., negation, hedging) in the context to inform the graph representations of knowledge. Our results on three benchmarks in the commonsense reasoning (i.e., CommonsenseQA, OpenbookQA) and medical question answering (i.e., MedQA-USMLE) domains demonstrate that GreaseLM can more reliably answer questions that require reasoning over both situational constraints and structured knowledge, even outperforming models 8x larger.

Learning meta-features for AutoML

- Herilalaina Rakotoarison, Louisot Milijaona, Andry RASOANAIVO, Michele Sebag, Marc Schoenauer
- abstract@open-review(Spotlight): This paper tackles the AutoML problem, aimed to automatically select an ML algorithm and its hyper-parameter configuration most appropriate to the dataset at hand. The proposed approach, MetaBu, learns new meta-features via an Optimal Transport procedure, aligning the manually designed \mf s with the space of distributions on the hyper-parameter configurations. MetaBu meta-features, learned once and for all, induce a topology on the set of datasets that is exploited to define a distribution of promising hyper-parameter configurations amenable to AutoML. Experiments on the OpenML CC-18 benchmark demonstrate that using MetaBu meta-features boosts the performance of state of the art AutoML systems, AutoSklearn (Feurer et al. 2015) and Probabilistic Matrix Factorization (Fusi et al. 2018). Furthermore, the inspection of MetaBu meta-features gives some hints into when an ML algorithm does well. Finally, the topology based on MetaBu meta-features enables to estimate the intrinsic dimensionality of the OpenML benchmark w.r.t. a given ML algorithm or pipeline. The source code is available at https://github.com/luxusg1/metabu.

Latent Variable Sequential Set Transformers for Joint Multi-Agent Motion Prediction

- Roger Girgis, Florian Golemo, Felipe Codevilla, Martin Weiss, Jim Aldon D'Souza, Samira Ebrahimi Kahou, Felix Heide, Christopher Pal
- abstract@open-review(Spotlight): Robust multi-agent trajectory prediction is essential for the safe control of robotic systems. A major challenge is to efficiently learn a representation that approximates the true joint distribution of contextual, social, and temporal information to enable planning. We propose Latent Variable Sequential Set Transformers which are encoder-decoder architectures that generate scene-consistent multi-agent trajectories. We refer to these architectures as â€œAutoBotsâ€ . The encoder is a stack of interleaved temporal and social multi-head self-attention (MHSA) modules which alternately perform equivariant processing across the temporal and social dimensions. The decoder employs learnable seed parameters in combination with temporal and social MHSA modules allowing it to perform inference over the entire future scene in a single forward pass efficiently. AutoBots can produce either the trajectory of one ego-agent or a distribution over the future trajectories for all agents in the scene. For the single-agent prediction case, our model achieves top results on the global nuScenes vehicle motion prediction leaderboard, and produces strong results on the Argoverse vehicle prediction challenge. In the multi-agent setting, we evaluate on the synthetic partition of TrajNet++ dataset to showcase the modelâ€™s socially-consistent predictions. We also demonstrate our model on general sequences of sets and provide illustrative experiments modelling the sequential structure of the multiple strokes that make up symbols in the Omniglot data. A distinguishing feature of AutoBots is that all models are trainable on a single desktop GPU (1080 Ti) in under 48h.

Understanding Latent Correlation-Based Multiview Learning and Self-Supervision: An Identifiability Perspective

- Qi Lyu, Xiao Fu, Weiran Wang, Songtao Lu
- abstract@open-review(Spotlight): Multiple views of data, both naturally acquired (e.g., image and audio) and artificially produced (e.g., via adding different noise to data samples), have proven useful in enhancing representation learning. Natural views are often handled by multiview analysis tools, e.g., (deep) canonical correlation analysis [(D)CCA], while the artificial ones are frequently used in self-supervised learning (SSL) paradigms, e.g., BYOL and Barlow Twins. Both types of approaches often involve learning neural feature extractors such that the embeddings of data exhibit high cross-view correlations. Although intuitive, the effectiveness of correlation-based neural embedding is mostly empirically validated. This work aims to understand latent correlation maximization-based deep multiview learning from a latent component identification viewpoint. An intuitive generative model of multiview data is adopted, where the views are different nonlinear mixtures of shared and private components. Since the shared components are view/distortion-invariant, representing the data using such components is believed to reveal the identity of the samples effectively and robustly. Under this model, latent correlation maximization is shown to guarantee the extraction of the shared components across views (up to certain ambiguities). In addition, it is further shown that the private information in each view can be provably disentangled from the shared using proper regularization design. A finite

sample analysis, which has been rare in nonlinear mixture identifiability study, is also presented. The theoretical results and newly designed regularization are tested on a series of tasks.

Deconstructing the Inductive Biases of Hamiltonian Neural Networks

- Nate Gruver, Marc Anton Finzi, Samuel Don Stanton, Andrew Gordon Wilson
- abstract@open-review(Spotlight): Physics-inspired neural networks (NNs), such as Hamiltonian or Lagrangian NNs, dramatically outperform other learned dynamics models by leveraging strong inductive biases. These models, however, are challenging to apply to many real world systems, such as those that donâ€™t conserve energy or contain contacts, a common setting for robotics and reinforcement learning. In this paper, we examine the inductive biases that make physics-inspired models successful in practice. We show that, contrary to conventional wisdom, the improved generalization of HNNs is the result of modeling acceleration directly and avoiding artificial complexity from the coordinate system, rather than symplectic structure or energy conservation. We show that by relaxing the inductive biases of these models, we can match or exceed performance on energy-conserving systems while dramatically improving performance on practical, non-conservative systems. We extend this approach to constructing transition models for common Mujoco environments, showing that our model can appropriately balance inductive biases with the flexibility required for model-based control.

Memorizing Transformers

- Yuhuai Wu, Markus Norman Rabe, DeLesley Hutchins, Christian Szegedy
- abstract@open-review(Spotlight): Language models typically need to be trained or finetuned in order to acquire new knowledge, which involves updating their weights.
  We instead envision language models that can simply read and memorize new data at inference time, thus acquiring new knowledge immediately. In this work, we extend language models with the ability to memorize the internal representations of past inputs. We demonstrate that an approximate kNN lookup into a non-differentiable memory of recent (key, value) pairs improves language modeling across various benchmarks and tasks, including generic webtext (C4), math papers (arXiv), books (PG-19), code (Github), as well as formal theorems (Isabelle). We show that the performance steadily improves when we increase the size of memory up to 262K tokens. On benchmarks including code and mathematics, we find that the model is capable of making use of newly defined functions and theorems during test time.

Learning-Augmented k-means Clustering

- Jon C. Ergun, Zhili Feng, Sandeep Silwal, David Woodruff, Samson Zhou
- abstract@open-review(Spotlight): k-means clustering is a well-studied problem due to its wide applicability. Unfortunately, there exist strong theoretical limits on the performance of any algorithm for the k-means problem on worst-case inputs. To overcome this barrier, we consider a scenario where ``advice'' is provided to help perform clustering. Specifically, we consider the k-means problem augmented with a predictor that, given any point, returns its cluster label in an approximately optimal clustering up to some, possibly adversarial, error. We present an algorithm whose performance improves along with the accuracy of the predictor, even though na\"{i}vely following the accurate predictor can still lead to a high clustering cost. Thus if the predictor is sufficiently accurate, we can retrieve a close to optimal clustering with nearly optimal runtime, breaking known computational barriers for algorithms that do not have access to such advice. We evaluate our algorithms on real datasets and show significant improvements in the quality of clustering.

On the Uncomputability of Partition Functions in Energy-Based Sequence Models

- Chu-Cheng Lin, Arya D. McCarthy
- abstract@open-review(Spotlight): In this paper, we argue that energy-based sequence models backed by expressive parametric families can result in uncomputable and inapproximable partition functions. Among other things, this makes model selection--and therefore learning model parameters--not only difficult, but generally *undecidable*. The reason is that there are no good deterministic or randomized estimates of partition functions. Specifically, we exhibit a pathological example where under common assumptions, *no* useful importance sampling estimates of the partition function can guarantee to have variance bounded below a rational number. As alternatives, we consider sequence model families whose partition functions are computable (if they exist), but at the cost of reduced expressiveness. Our theoretical results suggest that statistical procedures with asymptotic guarantees and sheer (but finite) amounts of compute are not the only things that make sequence modeling work; computability concerns must not be neglected as we consider more expressive model parametrizations.

Perceiver IO: A General Architecture for Structured Inputs & Outputs

- Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, Joao Carreira
- abstract@open-review(Spotlight): A central goal of machine learning is the development of systems that can solve many problems in as many data domains as possible. Current architectures, however, cannot be applied beyond a small set of stereotyped settings, as they bake in domain & task assumptions or scale poorly to large inputs or outputs. In this work, we propose Perceiver IO, a general-purpose architecture that handles data from arbitrary settings while scaling linearly with the size of inputs and outputs. Our model augments the Perceiver with a flexible querying mechanism that enables outputs of various sizes and semantics, doing away with the need for task-specific architecture engineering. The same architecture achieves strong results on tasks spanning natural language and visual understanding, multi-task and multi-modal reasoning, and StarCraft II. As highlights, Perceiver IO outperforms a Transformer-based BERT baseline on the GLUE language benchmark despite removing input tokenization and achieves state-of-the-art performance on Sintel optical flow estimation with no explicit mechanisms for multiscale correspondence.

DR3: Value-Based Deep Reinforcement Learning Requires Explicit Regularization

- Aviral Kumar, Rishabh Agarwal, Tengyu Ma, Aaron Courville, George Tucker, Sergey Levine
- abstract@open-review(Spotlight): Despite overparameterization, deep networks trained via supervised learning are surprisingly easy to optimize and exhibit excellent generalization. One hypothesis to explain this is that overparameterized deep networks enjoy the benefits of implicit regularization induced by stochastic gradient descent, which favors parsimonious solutions that generalize well on test inputs. It is reasonable to surmise that deep reinforcement learning (RL) methods could also benefit from this effect. In this paper, we discuss how the implicit regularization effect of SGD seen in supervised learning could in fact be harmful in the offline deep RL setting, leading to poor generalization and degenerate feature representations. Our theoretical analysis shows that when existing models of implicit regularization are applied to temporal difference learning, the resulting derived regularizer favors degenerate solutions with excessive aliasing, in stark contrast to the supervised learning case. We back up these findings empirically, showing that feature representations learned by a deep network value function trained via bootstrapping can indeed become degenerate, aliasing the representations for state-action pairs that appear on either side of the Bellman backup. To address this issue, we derive the form of this implicit regularizer and, inspired by this derivation, propose a simple and effective explicit regularizer, called DR3, that counteracts the undesirable effects of this implicit regularizer. When combined with existing offline RL methods, DR3 substantially improves performance and stability, alleviating unlearning in Atari 2600 games, D4RL domains and robotic manipulation from images.

MT3: Multi-Task Multitrack Music Transcription

- Joshua P Gardner, Ian Simon, Ethan Manilow, Curtis Hawthorne, Jesse Engel
- abstract@open-review(Spotlight): Automatic Music Transcription (AMT), inferring musical notes from raw audio, is a challenging task at the core of music understanding. Unlike Automatic Speech Recognition (ASR), which typically focuses on the words of a single speaker, AMT often requires transcribing multiple instruments simultaneously, all while preserving fine-scale pitch and timing information. Further, many AMT datasets are ``low-resource", as even expert musicians find music transcription difficult and time-consuming. Thus, prior work has focused on task-specific architectures, tailored to the individual instruments of each task. In this work, motivated by the promising results of sequence-to-sequence transfer learning for low-resource Natural Language Processing (NLP), we demonstrate that a general-purpose Transformer model can perform multi-task AMT, jointly transcribing arbitrary combinations of musical instruments across several transcription datasets. We show this unified training framework achieves high-quality transcription results across a range of datasets, dramatically improving performance for low-resource instruments (such as guitar), while preserving strong performance for abundant instruments (such as piano). Finally, by expanding the scope of AMT, we expose the need for more consistent evaluation metrics and better dataset alignment, and provide a strong baseline for this new direction of multi-task AMT.

Does your graph need a confidence boost? Convergent boosted smoothing on graphs with tabular node features

- Jiuhai Chen, Jonas Mueller, Vassilis N. Ioannidis, Soji Adeshina, Yangkun Wang, Tom Goldstein, David Wipf
- abstract@open-review(Spotlight): Many practical modeling tasks require making predictions using tabular data composed of heterogeneous feature types (e.g., text-based, categorical, continuous, etc.). In this setting boosted decision trees and related ensembling techniques generally dominate real-world applications involving iid training/test sets. However, when there are relations between samples and the iid assumption is no longer reasonable, it remains unclear how to incorporate these dependencies within existing boosting pipelines. To this end, we propose a generalized framework for combining boosted trees and more general model ensembling techniques, with graph propagation layers that share node/sample information across edges connecting related samples. And unlike previous efforts to integrate graph-based models with boosting, our approach is anchored to a principled meta loss function such that provable convergence can be guaranteed under relatively mild assumptions. Across a variety of benchmarks involving non-iid graph data with tabular node features, our framework achieves comparable or superior performance.

Geometric and Physical Quantities improve E(3) Equivariant Message Passing

- Johannes Brandstetter, Rob Hesselink, Elise van der Pol, Erik J Bekkers, Max Welling
- abstract@open-review(Spotlight): Including covariant information, such as position, force, velocity or spin is important in many tasks in computational physics and chemistry. We introduce Steerable E(3) Equivariant Graph Neural Networks (SEGNNs) that generalise equivariant graph networks, such that node and edge attributes are not restricted to invariant scalars, but can contain covariant information, such as vectors or tensors. Our model, composed of steerable MLPs, is able to incorporate geometric and physical information in both the message and update functions. Through the definition of steerable node attributes, the MLPs provide a new class of activation functions for general use with steerable feature fields. We discuss ours and related work through the lens of equivariant non-linear convolutions, which further allows us to pin-point the successful components of SEGNNs: non-linear message aggregation improves upon classic linear (steerable) point convolutions; steerable messages improve upon recent equivariant graph networks that send invariant messages. We demonstrate the effectiveness of our method on several tasks in computational physics and chemistry and provide extensive ablation studies.

SphereFace2: Binary Classification is All You Need for Deep Face Recognition

- Yandong Wen, Weiyang Liu, Adrian Weller, Bhiksha Raj, Rita Singh
- abstract@open-review(Spotlight): State-of-the-art deep face recognition methods are mostly trained with a softmax-based multi-class classification framework. Despite being popular and effective, these methods still have a few shortcomings that limit empirical performance. In this paper, we start by identifying the discrepancy between training and evaluation in the existing multi-class classification framework and then discuss the potential limitations caused by the "competitive" nature of softmax normalization. Motivated by these limitations, we propose a novel binary classification training framework, termed SphereFace2. In contrast to existing methods, SphereFace2 circumvents the softmax normalization, as well as the corresponding closed-set assumption. This effectively bridges the gap between training and evaluation, enabling the representations to be improved individually by each binary classification task. Besides designing a specific well-performing loss function, we summarize a few general principles for this "one-vs-all" binary classification framework so that it can outperform current competitive methods. Our experiments on popular benchmarks demonstrate that SphereFace2 can consistently outperform state-of-the-art deep face recognition methods.

Boosting Randomized Smoothing with Variance Reduced Classifiers

- MiklÃ³s Z. HorvÃ¡th, Mark Niklas Mueller, Marc Fischer, Martin Vechev
- abstract@open-review(Spotlight): Randomized Smoothing (RS) is a promising method for obtaining robustness certiï¬ cates by evaluating a base model under noise. In this work, we: (i) theoretically motivate why ensembles are a particularly suitable choice as base models for RS, and (ii) empirically conï¬ rm this choice, obtaining state-of-the-art results in multiple settings. The key insight of our work is that the reduced variance of ensembles over the perturbations introduced in RS leads to signiï¬ cantly more consistent classiï¬ cations for a given input. This, in turn, leads to substantially increased certiï¬ able radii for samples close to the decision boundary. Additionally, we introduce key optimizations which enable an up to 55-fold decrease in sample complexity of RS for predetermined radii, thus drastically reducing its computational overhead. Experimentally, we show that ensembles of only 3 to 10 classiï¬ ers consistently improve on their strongest constituting model with respect to their average certiï¬ ed radius (ACR) by 5% to 21% on both CIFAR10 and ImageNet, achieving a new state-of-the-art ACR of 0.86 and 1.11, respectively. We release all code and models required to reproduce our results at https://github.com/eth-sri/smoothing-ensembles.

SOSP: Efficiently Capturing Global Correlations by Second-Order Structured Pruning

- Manuel Nonnenmacher, Thomas Pfeil, Ingo Steinwart, David Reeb
- abstract@open-review(Spotlight): Pruning neural networks reduces inference time and memory costs. On standard hardware, these benefits will be especially prominent if coarse-grained structures, like feature maps, are pruned. We devise two novel saliency-based methods for second-order structured pruning (SOSP) which include correlations among all structures and layers. Our main method SOSP-H employs an innovative second-order approximation, which enables saliency evaluations by fast Hessian-vector products. SOSP-H thereby scales like a first-order method despite taking into account the full Hessian. We validate SOSP-H by comparing it to our second method SOSP-I that uses a well-established Hessian approximation, and to numerous state-of-the-art methods. While SOSP-H performs on par or better in terms of accuracy, it has clear advantages in terms of scalability and efficiency. This allowed us to scale SOSP-H to large-scale vision tasks, even though it captures correlations across all layers of the network. To underscore the global nature of our pruning methods, we evaluate their performance not only by removing structures from a pretrained network, but also by detecting architectural bottlenecks. We show that our algorithms allow to systematically reveal architectural bottlenecks, which we then remove to further increase the accuracy of the networks.

Relational Multi-Task Learning: Modeling Relations between Data and Tasks

- Kaidi Cao, Jiaxuan You, Jure Leskovec

- abstract@open-review(Spotlight): A key assumption in multi-task learning is that at the inference time the multi-task model only has access to a given data point but not to the data pointâ€™s labels from other tasks. This presents an opportunity to extend multi-task learning to utilize data pointâ€™s labels from other auxiliary tasks, and this way improves performance on the new task. Here we introduce a novel relational multi-task learning setting where we leverage data point labels from auxiliary tasks to make more accurate predictions on the new task. We develop MetaLink, where our key innovation is to build a knowledge graph that connects data points and tasks and thus allows us to leverage labels from auxiliary tasks. The knowledge graph consists of two types of nodes: (1) data nodes, where node features are data embeddings computed by the neural network, and (2) task nodes, with the last layerâ€™s weights for each task as node features. The edges in this knowledge graph capture data-task relationships, and the edge label captures the label of a data point on a particular task. Under MetaLink, we reformulate the new task as a link label prediction problem between a data node and a task node. The MetaLink framework provides flexibility to model knowledge transfer from auxiliary task labels to the task of interest. We evaluate MetaLink on 6 benchmark datasets in both biochemical and vision domains. Experiments demonstrate that MetaLink can successfully utilize the relations among different tasks, outperforming the state-of-the-art methods under the proposed relational multi-task learning setting, with up to 27% improvement in ROC AUC.

CoBERL: Contrastive BERT for Reinforcement Learning

- Andrea Banino, Adria Puigdomenech Badia, Jacob C Walker, Tim Scholtes, Jovana Mitrovic, Charles Blundell
- abstract@open-review(Spotlight): Many reinforcement learning (RL) agents require a large amount of experience to solve tasks. We propose Contrastive BERT for RL (COBERL), an agent that combines a new contrastive loss and a hybrid LSTM-transformer architecture to tackle the challenge of improving data efficiency. COBERL enables efficient and robust learning from pixels across a wide variety of domains. We use bidirectional masked prediction in combination with a generalization of a recent contrastive method to learn better representations for RL, without the need of hand engineered data augmentations. We find that COBERL consistently improves data efficiency across the full Atari suite, a set of control tasks and a challenging 3D environment, and often it also increases final score performance.

Optimal Transport for Causal Discovery

- Ruibo Tu, Kun Zhang, Hedvig Kjellstrom, Cheng Zhang
- abstract@open-review(Spotlight): To determine causal relationships between two variables, approaches based on Functional Causal Models (FCMs) have been proposed by properly restricting model classes; however, the performance is sensitive to the model assumptions, which makes it difficult to use. In this paper, we provide a novel dynamical-system view of FCMs and propose a new framework for identifying causal direction in the bivariate case. We first show the connection between FCMs and optimal transport, and then study optimal transport under the constraints of FCMs. Furthermore, by exploiting the dynamical interpretation of optimal transport under the FCM constraints, we determine the corresponding underlying dynamical process of the static cause-effect pair data. It provides a new dimension for describing static causal discovery tasks while enjoying more freedom for modeling the quantitative causal influences. In particular, we show that Additive Noise Models (ANMs) correspond to volume-preserving pressureless flows. Consequently, based on their velocity field divergence, we introduce a criterion for determining causal direction. With this criterion, we propose a novel optimal transport-based algorithm for ANMs which is robust to the choice of models and extend it to post-nonlinear models. Our method demonstrated state-of-the-art results on both synthetic and causal discovery benchmark datasets.

On Bridging Generic and Personalized Federated Learning for Image Classification

- Hong-You Chen, Wei-Lun Chao
- abstract@open-review(Spotlight): Federated learning is promising for its capability to collaboratively train models with multiple clients without accessing their data, but vulnerable when clients' data distributions diverge from each other. This divergence further leads to a dilemma: "Should we prioritize the learned model's generic performance (for future use at the server) or its personalized performance (for each client)?" These two, seemingly competing goals have divided the community to focus on one or the other, yet in this paper we show that it is possible to approach both at the same time. Concretely, we propose a novel federated learning framework that explicitly decouples a model's dual duties with two prediction tasks. On the one hand, we introduce a family of losses that are robust to non-identical class distributions, enabling clients to train a generic predictor with a consistent objective across them. On the other hand, we formulate the personalized predictor as a lightweight adaptive module that is learned to minimize each client's empirical risk on top of the generic predictor. With this two-loss, two-predictor framework which we name Federated Robust Decoupling (Fed-RoD), the learned model can simultaneously achieve state-of-the-art generic and personalized performance, essentially bridging the two tasks.

Value Gradient weighted Model-Based Reinforcement Learning

- Claas A Voelcker, Victor Liao, Animesh Garg, Amir-massoud Farahmand
- abstract@open-review(Spotlight): Model-based reinforcement learning (MBRL) is a sample efficient technique to obtain control policies, yet unavoidable modeling errors often lead performance deterioration. The model in MBRL is often solely fitted to reconstruct dynamics, state observations in particular, while the impact of model error on the policy is not captured by the training objective. This leads to a mismatch between the intended goal of MBRL, enabling good policy and value learning, and the target of the loss function employed in practice, future state prediction. Naive intuition would suggest that value-aware model learning would fix this problem and, indeed, several solutions to this objective mismatch problem have been proposed based on theoretical analysis. However, they tend to be inferior in practice to commonly used maximum likelihood (MLE) based approaches. In this paper we propose the Value-gradient weighted Model Learning (VaGraM), a novel method for value-aware model learning which improves the performance of MBRL in challenging settings, such as small model capacity and the presence of distracting state dimensions. We analyze both MLE and value-aware approaches and demonstrate how they fail to account for exploration and the behavior of function approximation when learning value-aware models and highlight the additional goals that must be met to stabilize optimization in the deep learning setting. We verify our analysis by showing that our loss function is able to achieve high returns on the Mujoco benchmark suite while being more robust than maximum likelihood based approaches.

Fairness in Representation for Multilingual NLP: Insights from Controlled Experiments on Conditional Language Modeling

- Ada Wan
- abstract@open-review(Spotlight): We perform systematically and fairly controlled experiments with the 6-layer Transformer to investigate the hardness in conditional-language-modeling languages which have been traditionally considered morphologically rich (AR and RU) and poor (ZH). We evaluate through statistical comparisons across 30 possible language directions from the 6 languages of the United Nations Parallel Corpus across 5 data sizes on 3 representation levels --- character, byte, and word. Results show that performance is relative to the representation granularity of each of the languages, not to the language as a whole. On the character and byte levels, we are able to eliminate statistically significant performance disparity, hence demonstrating that a language cannot be intrinsically hard. The disparity that mirrors the morphological complexity hierarchy is shown to be a byproduct of word segmentation. Evidence from data statistics, along with the fact that word segmentation is qualitatively indeterminate, renders a decades-long debate on morphological complexity (unless it is being intentionally modeled in a word-based, meaning-driven context) irrelevant in the context of computing. The intent of our work is to help effect more objectivity and adequacy in evaluation as well as fairness and inclusivity in experimental setup in the area of language and computing so to uphold diversity in Machine Learning and Artificial Intelligence research. Multilinguality is real and relevant in computing not due to canonical, structural linguistic concepts such as morphology or "words" in our minds, but rather standards related to internationalization and localization, such as character encoding --- something which has thus far been sorely overlooked in our discourse and curricula.

Reinforcement Learning with Sparse Rewards using Guidance from Offline Demonstration

- Desik Rengarajan, Gargi Vaidya, Akshay Sarvesh, Dileep Kalathil, Srinivas Shakkottai
- abstract@open-review(Spotlight): A major challenge in real-world reinforcement learning (RL) is the sparsity of reward feedback. Often, what is available is an intuitive but sparse reward function that only indicates whether the task is completed partially or fully. However, the lack of carefully designed, fine grain feedback implies that most existing RL algorithms fail to learn an acceptable policy in a reasonable time frame. This is because of the large number of exploration actions that the policy has to perform before it gets any useful feedback that it can learn from. In this work, we address this challenging problem by developing an algorithm that exploits the offline demonstration data generated by {a sub-optimal behavior policy} for faster and efficient online RL in such sparse reward settings. The proposed algorithm, which we call the Learning Online with Guidance Offline (LOGO) algorithm, merges a policy improvement step with an additional policy guidance step by using the offline demonstration data. The key idea is that by obtaining guidance from - not imitating - the offline {data}, LOGO orients its policy in the manner of the sub-optimal {policy}, while yet being able to learn beyond and approach optimality. We provide a theoretical analysis of our algorithm, and provide a lower bound on the performance improvement in each learning episode. We also extend our algorithm to the even more challenging incomplete observation setting, where the demonstration data contains only a censored version of the true state observation. We demonstrate the superior performance of our algorithm over state-of-the-art approaches on a number of benchmark environments with sparse rewards {and censored state}. Further, we demonstrate the value of our approach via implementing LOGO on a mobile robot for trajectory tracking and obstacle avoidance, where it shows excellent performance.

Linking Emergent and Natural Languages via Corpus Transfer

- Shunyu Yao, Mo Yu, Yang Zhang, Karthik R Narasimhan, Joshua B. Tenenbaum, Chuang Gan
- abstract@open-review(Spotlight): The study of language emergence aims to understand how human languages are shaped by perceptual grounding and communicative intent. Computational approaches to emergent communication (EC) predominantly consider referential games in limited domains and analyze the learned protocol within the game framework. As a result, it remains unclear how the emergent languages from these settings connect to natural languages or provide benefits in real-world language processing tasks, where statistical models trained on large text corpora dominate. In this work, we propose a novel way to establish such a link by corpus transfer, i.e. pretraining on a corpus of emergent language for downstream natural language tasks, which is in contrast to prior work that directly transfers speaker and listener parameters. Our approach showcases non-trivial transfer benefits for two different tasks â€“ language modeling and image captioning. For example, in a low-resource setup (modeling 2 million natural language tokens), pre-training on an emergent language corpus with just 2 million tokens reduces model perplexity by 24.6% on average across ten natural languages. We also introduce a novel metric to predict the transferability of an emergent language by translating emergent messages to natural language captions grounded on the same images. We find that our translation-based metric highly correlates with the downstream performance on modeling natural languages (for instance $\rho = 0.83$ on Hebrew), while topographic similarity, a popular metric in previous works, shows surprisingly low correlation ($\rho = 0.003$), hinting that simple properties like attribute disentanglement from synthetic domains might not capture the full complexities of natural language. Our findings also indicate potential benefits of moving language emergence forward with natural language resources and models.

TAMP-S2GCNets: Coupling Time-Aware Multipersistence Knowledge Representation with Spatio-Supra Graph Convolutional Networks for Time-Series Forecasting

- Yuzhou Chen, Ignacio Segovia-Dominguez, Baris Coskunuzer, Yulia Gel
- abstract@open-review(Spotlight): Graph Neural Networks (GNNs) are proven to be a powerful machinery for learning complex dependencies in multivariate spatio-temporal processes. However, most existing GNNs have inherently static architectures, and as a result, do not explicitly account for time dependencies of the encoded knowledge and are limited in their ability to simultaneously infer latent time-conditioned relations among entities. We postulate that such hidden time-conditioned properties may be captured by the tools of multipersistence, i.e, a emerging machinery in topological data analysis which allows us to quantify dynamics of the data shape along multiple geometric dimensions. We make the first step toward integrating the two rising research directions, that is, time-aware deep learning and multipersistence, and propose a new model, Time-Aware Multipersistence Spatio-Supra Graph Convolutional Network (TAMP-S2GCNets). We summarize inherent time-conditioned topological properties of the data as time-aware multipersistence Euler-Poincar\'e surface and prove its stability. We then construct a supragraph convolution module which simultaneously accounts for the extracted intra- and inter- spatio-temporal dependencies in the data. Our extensive experiments on highway traffic flow, Ethereum token prices, and COVID-19 hospitalizations demonstrate that TAMP-S2GCNets outperforms the state-of-the-art tools in multivariate time series forecasting tasks.

The MultiBERTs: BERT Reproductions for Robustness Analysis

- Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Raluca Turc, Jacob Eisenstein, Dipanjan Das, Ellie Pavlick
- abstract@open-review(Spotlight): Experiments with pre-trained models such as BERT are often based on a single checkpoint. While the conclusions drawn apply to the artifact tested in the experiment (i.e., the particular instance of the model), it is not always clear whether they hold for the more general procedure which includes the architecture, training data, initialization scheme, and loss function. Recent work has shown that repeating the pre-training process can lead to substantially different performance, suggesting that an alternative strategy is needed to make principled statements about procedures. To enable researchers to draw more robust conclusions, we introduce MultiBERTs, a set of 25 BERT-Base checkpoints, trained with similar hyper-parameters as the original BERT model but differing in random weight initialization and shuffling of training data. We also define the Multi-Bootstrap, a non-parametric bootstrap method for statistical inference designed for settings where there are multiple pre-trained models and limited test data. To illustrate our approach, we present a case study of gender bias in coreference resolution, in which the Multi-Bootstrap lets us measure effects that may not be detected with a single checkpoint. The models and statistical library are available online, along with an additional set of 140 intermediate checkpoints captured during pre-training to facilitate research on learning dynamics.

Message Passing Neural PDE Solvers

- Johannes Brandstetter, Daniel E. Worrall, Max Welling
- abstract@open-review(Spotlight): The numerical solution of partial differential equations (PDEs) is difficult, having led to a century of research so far. Recently, there have been pushes to build neural--numerical hybrid solvers, which piggy-backs the modern trend towards fully end-to-end learned systems. Most works so far can only generalize over a subset of properties to which a generic solver would be faced, including: resolution, topology, geometry, boundary conditions, domain discretization regularity, dimensionality, etc. In this work, we build a solver, satisfying these properties, where all the components are based on neural message passing, replacing all heuristically designed components in the computation graph with backprop-optimized neural function approximators. We show that neural message passing solvers representationally contain some classical methods, such as finite differences, finite volumes, and WENO schemes. In order to encourage stability in training autoregressive models, we put forward a method that is based on the principle of zero-stability, posing stability as a domain adaptation problem. We validate our method on various fluid-like flow problems, demonstrating fast, stable, and accurate performance across different domain topologies, discretization, etc. in 1D and 2D. Our model outperforms state-of-the-art numerical solvers in the low resolution regime in terms of speed, and accuracy.

Multi-Stage Episodic Control for Strategic Exploration in Text Games

- Jens Tuyls, Shunyu Yao, Sham M. Kakade, Karthik R Narasimhan
- abstract@open-review(Spotlight): Text adventure games present unique challenges to reinforcement learning methods due to their combinatorially large action spaces and sparse rewards. The interplay of these two factors is particularly demanding because large action spaces require extensive exploration, while sparse rewards provide limited feedback. This work proposes to tackle the explore-vs-exploit dilemma using a multi-stage approach that explicitly

disentangles these two strategies within each episode. Our algorithm, called eXploit-Then-eXplore (XTX), begins each episode using an exploitation policy that imitates a set of promising trajectories from the past, and then switches over to an exploration policy aimed at discovering novel actions that lead to unseen state spaces. This policy decomposition allows us to combine global decisions about which parts of the game space to return to with curiosity-based local exploration in that space, motivated by how a human may approach these games. Our method significantly outperforms prior approaches by 27% and 11% average normalized score over 12 games from the Jericho benchmark (Hausknecht et al., 2020) in both deterministic and stochastic settings, respectively. On the game of Zork1, in particular, XTX obtains a score of 103, more than a 2x improvement over prior methods, and pushes past several known bottlenecks in the game that have plagued previous state-of-the-art methods.

Exploring the Limits of Large Scale Pre-training

- Samira Abnar, Mostafa Dehghani, Behnam Neyshabur, Hanie Sedghi
- abstract@open-review(Spotlight): Recent developments in large-scale machine learning suggest that by scaling up data, model size and training time properly, one might observe that improvements in pre-training would transfer favorably to most downstream tasks. In this work we systematically study this phenomena and establish that, as we increase the upstream accuracy, performance of downstream tasks \emph{saturates}. In particular, we investigate more than 4800 experiments on Vision Transformers, MLP-Mixers and ResNets with number of parameters ranging from ten million to ten billion, trained on the largest scale of available image data (JFT, ImageNet21K) and evaluated on more than 20 downstream image recognition tasks. We propose a model for downstream performance that reflects the saturation phenomena and captures the nonlinear relationship in performance of upstream and downstream tasks. Delving deeper to understand the reasons that give rise to these phenomena, we show that the observed saturation behavior is closely related to the way that representations evolve through the layers of the models. We showcase an even more extreme scenario where performance on upstream and downstream are at odds with each other. That is, in order to have a better downstream performance, we need to hurt upstream accuracy.

Universal Approximation Under Constraints is Possible with Transformers

- Anastasis Kratsios, Behnoosh Zamanlooy, Tianlin Liu, Ivan DokmaniÄ‡
- abstract@open-review(Spotlight): Many practical problems need the output of a machine learning model to satisfy a set of constraints, K. Nevertheless, there is no known guarantee that classical neural network architectures can exactly encode constraints while simultaneously achieving universality. We provide a quantitative constrained universal approximation theorem which guarantees that for any non-convex compact set K and any continuous function $f:\mathbb{R}^n\rightarrow K$, there is a probabilistic transformer $\hat{F}$ whose randomized outputs all lie in K and whose expected output uniformly approximates f. Our second main result is a ``deep neural version" of Berge's Maximum Theorem (1963). The result guarantees that given an objective function L, a constraint set K, and a family of soft constraint sets, there is a probabilistic transformer $\hat{F}$ that approximately minimizes L and whose outputs belong to K; moreover, $\hat{F}$ approximately satisfies the soft constraints. Our results imply the first universal approximation theorem for classical transformers with exact convex constraint satisfaction. They also yield that a chart-free universal approximation theorem for Riemannian manifold-valued functions subject to suitable geodesically convex constraints.

Scaling Laws for Neural Machine Translation

- Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, Colin Cherry
- abstract@open-review(Spotlight): We present an empirical study of scaling properties of encoder-decoder Transformer models used in neural machine translation (NMT). We show that cross-entropy loss as a function of model size follows a certain scaling law. Specifically (i) We propose a formula which describes the scaling behavior of cross-entropy loss as a bivariate function of encoder and decoder size, and show that it gives accurate predictions under a variety of scaling approaches and languages; we show that the total number of parameters alone is not sufficient for such purposes. (ii) We observe different power law exponents when scaling the decoder vs scaling the encoder, and provide recommendations for optimal allocation of encoder/decoder capacity based on this observation. (iii) We also report that the scaling behavior of the model is acutely influenced by composition bias of the train/test sets, which we define as any deviation from naturally generated text (either via machine generated or human translated text). We observe that natural text on the target side enjoys scaling, which manifests as successful reduction of the cross-entropy loss. (iv) Finally, we investigate the relationship between the cross-entropy loss and the quality of the generated translations. We find two different behaviors, depending on the nature of the test data. For test sets which were originally translated from target language to source language, both loss and BLEU score improve as model size increases. In contrast, for test sets originally translated from source language to target language, the loss improves, but the BLEU score stops improving after a certain threshold. We release generated text from all models used in this study.

AdaRL: What, Where, and How to Adapt in Transfer Reinforcement Learning

- Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, Kun Zhang
- abstract@open-review(Spotlight): One practical challenge in reinforcement learning (RL) is how to make quick adaptations when faced with new environments. In this paper, we propose a principled framework for adaptive RL, called AdaRL, that adapts reliably and efficiently to changes across domains with a few samples from the target domain, even in partially observable environments. Specifically, we leverage a parsimonious graphical representation that characterizes structural relationships over variables in the RL system. Such graphical representations provide a compact way to encode what and where the changes across domains are, and furthermore inform us with a minimal set of changes that one has to consider for the purpose of policy adaptation. We show that by explicitly leveraging this compact representation to encode changes, we can efficiently adapt the policy to the target domain, in which only a few samples are needed and further policy optimization is avoided. We illustrate the efficacy of AdaRL through a series of experiments that vary factors in the observation, transition and reward functions for Cartpole and Atari games.

Independent SE(3)-Equivariant Models for End-to-End Rigid Protein Docking

- Octavian-Eugen Ganea, Xinyuan Huang, Charlotte Bunne, Yatao Bian, Regina Barzilay, Tommi S. Jaakkola, Andreas Krause
- abstract@open-review(Spotlight): Protein complex formation is a central problem in biology, being involved in most of the cell's processes, and essential for applications, e.g. drug design or protein engineering. We tackle rigid body protein-protein docking, i.e., computationally predicting the 3D structure of a protein-protein complex from the individual unbound structures, assuming no conformational change within the proteins happens during binding. We design a novel pairwise-independent SE(3)-equivariant graph matching network to predict the rotation and translation to place one of the proteins at the right docked position relative to the second protein. We mathematically guarantee a basic principle: the predicted complex is always identical regardless of the initial locations and orientations of the two structures. Our model, named EquiDock, approximates the binding pockets and predicts the docking poses using keypoint matching and alignment, achieved through optimal transport and a differentiable Kabsch algorithm. Empirically, we achieve significant running time improvements and often outperform existing docking software despite not relying on heavy candidate sampling, structure refinement, or templates.

Towards a Unified View of Parameter-Efficient Transfer Learning

- Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, Graham Neubig
- abstract@open-review(Spotlight): Fine-tuning large pretrained language models on downstream tasks has become the de-facto learning paradigm in NLP. However, conventional approaches fine-tune all the parameters of the pretrained model, which becomes prohibitive as the model size and the number of tasks grow. Recent work has proposed a variety of parameter-efficient transfer learning methods that only fine-tune a small number of (extra) parameters to attain strong performance. While effective, the critical ingredients for success and the connections among the various methods are poorly understood. In

this paper, we break down the design of state-of-the-art parameter-efficient transfer learning methods and present a unified framework that establishes connections between them. Specifically, we re-frame them as modifications to specific hidden states in pretrained models, and define a set of design dimensions along which different methods vary, such as the function to compute the modification and the position to apply the modification. Through comprehensive empirical studies across machine translation, text summarization, language understanding, and text classification benchmarks, we utilize the unified view to identify important design choices in previous methods. Furthermore, our unified framework enables the transfer of design elements across different approaches, and as a result we are able to instantiate new parameter-efficient fine-tuning methods that tune less parameters than previous methods while being more effective, achieving comparable results to fine-tuning all parameters on all four tasks.

GNN-LM: Language Modeling based on Global Contexts via GNN

- Yuxian Meng, Shi Zong, Xiaoya Li, Xiaofei Sun, Tianwei Zhang, Fei Wu, Jiwei Li
- abstract@open-review(Spotlight): Inspired by the notion that "it to copy is easier than to memorize", in this work, we introduce GNN-LM, which extends vanilla neural language model (LM) by allowing to reference similar contexts in the entire training corpus. We build a directed heterogeneous graph between an input context and its semantically related neighbors selected from the training corpus, where nodes are tokens in the input context and retrieved neighbor contexts, and edges represent connections between nodes. Graph neural networks (GNNs) are constructed upon the graph to aggregate information from similar contexts to decode the token. This learning paradigm provides direct access to the reference contexts and helps improve a model's generalization ability. We conduct comprehensive experiments to validate the effectiveness of the GNN-LM: GNN-LM achieves a new state-of-the-art perplexity of 14.8 on WikiText-103 (a 3.9 point improvement over its counterpart of the vanilla LM model), and shows substantial improvement on One Billion Word and Enwiki8 datasets against strong baselines. In-depth ablation studies are performed to understand the mechanics of GNN-LM. The code can be found at https://github.com/ShannonAI/GNN-LM.

Continual Learning with Filter Atom Swapping

- Zichen Miao, Ze Wang, Wei Chen, Qiang Qiu
- abstract@open-review(Spotlight): Continual learning has been widely studied in recent years to resolve the catastrophic forgetting of deep neural networks. In this paper, we first enforce a low-rank filter subspace by decomposing convolutional filters within each network layer over a small set of filter atoms. Then, we perform continual learning with filter atom swapping. In other words, we learn for each task a new filter subspace for each convolutional layer, i.e., hundreds of parameters as filter atoms, but keep subspace coefficients shared across tasks. By maintaining a small footprint memory of filter atoms, we can easily archive models for past tasks to avoid forgetting. The effectiveness of this simple scheme for continual learning is illustrated both empirically and theoretically. The proposed atom swapping framework further enables flexible and efficient model ensemble with members selected within a task or across tasks to improve the performance in different continual learning settings. Being validated on multiple benchmark datasets with different convolutional network structures, the proposed method outperforms the state-of-the-art methods in both accuracy and scalability.

Continual Learning with Recursive Gradient Optimization

- Hao Liu, Huaping Liu
- abstract@open-review(Spotlight): Learning multiple tasks sequentially without forgetting previous knowledge, called Continual Learning(CL), remains a long-standing challenge for neural networks. Most existing methods rely on additional network capacity or data replay. In contrast, we introduce a novel approach which we refer to as Recursive Gradient Optimization(RGO). RGO is composed of an iteratively updated optimizer that modifies the gradient to minimize forgetting without data replay and a virtual Feature Encoding Layer(FEL) that represents different long-term structures with only task descriptors. Experiments demonstrate that RGO has significantly better performance on popular continual classification benchmarks when compared to the baselines and achieves new state-of-the-art performance on 20-split-CIFAR100(82.22%) and 20-split-miniImageNet(72.63%). With higher average accuracy than Single-Task Learning(STL), this method is flexible and reliable to provide continual learning capabilities for learning models that rely on gradient descent.

NODE-GAM: Neural Generalized Additive Model for Interpretable Deep Learning

- Chun-Hao Chang, Rich Caruana, Anna Goldenberg
- abstract@open-review(Spotlight): Deployment of machine learning models in real high-risk settings (e.g. healthcare) often depends not only on the model's accuracy but also on its fairness, robustness, and interpretability. Generalized Additive Models (GAMs) are a class of interpretable models with a long history of use in these high-risk domains, but they lack desirable features of deep learning such as differentiability and scalability. In this work, we propose a neural GAM (NODE-GAM) and neural GA^{2}M (NODE-GA^{2}M) that scale well and perform better than other GAMs on large datasets, while remaining interpretable compared to other ensemble and deep learning models. We demonstrate that our models find interesting patterns in the data. Lastly, we show that we are able to improve model accuracy via self-supervised pre-training, an improvement that is not possible for non-differentiable GAMs.

Learnability of convolutional neural networks for infinite dimensional input via mixed and anisotropic smoothness

- Sho Okumoto, Taiji Suzuki
- abstract@open-review(Spotlight): Among a wide range of success of deep learning, convolutional neural networks have been extensively utilized in several tasks such as speech recognition, image processing, and natural language processing, which require inputs with large dimensions. Several studies have investigated function estimation capability of deep learning, but most of them have assumed that the dimensionality of the input is much smaller than the sample size. However, for typical data in applications such as those handled by the convolutional neural networks described above, the dimensionality of inputs is relatively high or even infinite. In this paper, we investigate the approximation and estimation errors of the (dilated) convolutional neural networks when the input is infinite dimensional. Although the approximation and estimation errors of neural networks are affected by the curse of dimensionality in the existing analyses for typical function spaces such as the \Holder and Besov spaces, we show that, by considering anisotropic smoothness, they can alleviate exponential dependency on the dimensionality but they only depend on the smoothness of the target functions. Our theoretical analysis supports the great practical success of convolutional networks.
  Furthermore, we show that the dilated convolution is advantageous when the smoothness of the target function has a sparse structure.

Improved deterministic l2 robustness on CIFAR-10 and CIFAR-100

- Sahil Singla, Surbhi Singla, Soheil Feizi
- abstract@open-review(Spotlight): Training convolutional neural networks (CNNs) with a strict Lipschitz constraint under the l_{2} norm is useful for provable adversarial robustness, interpretable gradients and stable training. While 1-Lipschitz CNNs can be designed by enforcing a 1-Lipschitz constraint on each layer, training such networks requires each layer to have an orthogonal Jacobian matrix (for all inputs) to prevent the gradients from vanishing during backpropagation. A layer with this property is said to be Gradient Norm Preserving (GNP). In this work, we introduce a procedure to certify the robustness of 1-Lipschitz CNNs by relaxing the orthogonalization of the last linear layer of the network that significantly advances the state of the art for both standard and provable robust accuracies on CIFAR-100 (gains of 4.80% and 4.71%, respectively). We further boost their robustness by introducing (i) a novel Gradient Norm preserving activation function called the Householder activation function (that includes every $\mathrm{GroupSort}$ activation) and (ii) a certificate regularization. On CIFAR-10, we achieve significant improvements over prior works in provable

robust accuracy (5.81%) with only a minor drop in standard accuracy (-0.29%). Code for reproducing all experiments in the paper is available at \url{https://github.com/singlasahil14/SOC}.

Transition to Linearity of Wide Neural Networks is an Emerging Property of Assembling Weak Models

- Chaoyue Liu, Libin Zhu, Misha Belkin
- abstract@open-review(Spotlight): Wide neural networks with linear output layer have been shown to be near-linear, and to have near-constant neural tangent kernel (NTK), in a region containing the optimization path of gradient descent. These findings seem counter-intuitive since in general neural networks are highly complex models. Why does a linear structure emerge when the neural networks become wide? In this work, we provide a new perspective on this "transition to linearity" by considering a neural network as an assembly model recursively built from a set of sub-models corresponding to individual neurons. In this view, we show that the linearity of wide neural networks is, in fact, an emerging property of assembling a large number of diverse ``weak'' sub-models, none of which dominate the assembly.

Looking Back on Learned Experiences For Class/task Incremental Learning

- Mozhgan PourKeshavarzi, Guoying Zhao, Mohammad Sabokrou
- abstract@open-review(Spotlight): Classical deep neural networks are limited in their ability to learn from emerging streams of training data. When trained sequentially on new or evolving tasks, their performance degrades sharply, making them inappropriate in real-world use cases. Existing methods tackle it by either storing old data samples or only updating a parameter set of deep neural networks, which, however, demands a large memory budget or spoils the flexibility of models to learn the incremented task distribution. In this paper, we shed light on an on-call transfer set to provide past experiences whenever a new task arises in the data stream. In particular, we propose a Cost-Free Incremental Learning (CF-IL) not only to replay past experiences the model has learned but also to perform this in a cost free manner. Towards this end, we introduced a memory recovery paradigm in which we query the network to synthesize past exemplars whenever a new task emerges. Thus, our method needs no extra memory for data buffering or network growing, besides calls the proposed memory recovery paradigm to provide past exemplars, named a transfer set in order to mitigate catastrophically forgetting the former tasks in the Incremental Learning (IL) setup. Moreover, in contrast with recently proposed methods, the suggested paradigm does not desire a parallel architecture since it only relies on the learner network. Compared to the state-of-the-art data techniques without buffering past data samples, CF-IL demonstrates significantly better performance on the well-known datasets whether a task oracle is available in test time (Task-IL) or not (Class-IL).

EntQA: Entity Linking as Question Answering

- Wenzheng Zhang, Wenyue Hua, Karl Stratos
- abstract@open-review(Spotlight): A conventional approach to entity linking is to first find mentions in a given document and then infer their underlying entities in the knowledge base. A well-known limitation of this approach is that it requires finding mentions without knowing their entities, which is unnatural and difficult. We present a new model that does not suffer from this limitation called $\textbf{EntQA}$, which stands for $\mbox{\textbf{Ent}ity}$ linking as $\mbox{\textbf{Q}uestion}$ $\mbox{\textbf{A}nswering}$. EntQA first proposes candidate entities with a fast retrieval module, and then scrutinizes the document to find mentions of each candidate with a powerful reader module. Our approach combines progress in entity linking with that in open-domain question answering and capitalizes on pretrained models for dense entity retrieval and reading comprehension. Unlike in previous works, we do not rely on a mention-candidates dictionary or large-scale weak supervision. EntQA achieves strong results on the GERBIL benchmarking platform.

Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems

- Thomas Pethick, Puya Latafat, Panos Patrinos, Olivier Fercoq, Volkan Cevher
- abstract@open-review(Spotlight): This paper introduces a new extragradient-type algorithm for a class of nonconvex-nonconcave minimax problems. It is well-known that finding a local solution for general minimax problems is computationally intractable. This observation has recently motivated the study of structures sufficient for convergence of first order methods in the more general setting of variational inequalities when the so-called weak Minty variational inequality (MVI) holds. This problem class captures non-trivial structures as we demonstrate with examples, for which a large family of existing algorithms provably converge to limit cycles. Our results require a less restrictive parameter range in the weak MVI compared to what is previously known, thus extending the applicability of our scheme. The proposed algorithm is applicable to constrained and regularized problems, and involves an adaptive stepsize allowing for potentially larger stepsizes. Our scheme also converges globally even in settings where the underlying operator exhibits limit cycles.

Compositional Attention: Disentangling Search and Retrieval

- Sarthak Mittal, Sharath Chandra Raparthy, Irina Rish, Yoshua Bengio, Guillaume Lajoie
- abstract@open-review(Spotlight): Multi-head, key-value attention is the backbone of transformer-like model architectures which have proven to be widely successful in recent years. This attention mechanism uses multiple parallel key-value attention blocks (called heads), each performing two fundamental computations: (1) search - selection of a relevant entity from a set via query-key interaction, and (2) retrieval - extraction of relevant features from the selected entity via a value matrix. Standard attention heads learn a rigid mapping between search and retrieval. In this work, we first highlight how this static nature of the pairing can potentially: (a) lead to learning of redundant parameters in certain tasks, and (b) hinder generalization. To alleviate this problem, we propose a novel attention mechanism, called Compositional Attention, that replaces the standard head structure. The proposed mechanism disentangles search and retrieval and composes them in a dynamic, flexible and context-dependent manner. Through a series of numerical experiments, we show that it outperforms standard multi-head attention on a variety of tasks, including some out-of-distribution settings. Through our qualitative analysis, we demonstrate that Compositional Attention leads to dynamic specialization based on the type of retrieval needed. Our proposed mechanism generalizes multi-head attention, allows independent scaling of search and retrieval and is easy to implement in a variety of established network architectures.

Contrastive Fine-grained Class Clustering via Generative Adversarial Networks

- Yunji Kim, Jung-Woo Ha
- abstract@open-review(Spotlight): Unsupervised fine-grained class clustering is a practical yet challenging task due to the difficulty of feature representations learning of subtle object details. We introduce C3-GAN, a method that leverages the categorical inference power of InfoGAN with contrastive learning. We aim to learn feature representations that encourage a dataset to form distinct cluster boundaries in the embedding space, while also maximizing the mutual information between the latent code and its image observation. Our approach is to train a discriminator, which is also used for inferring clusters, to optimize the contrastive loss, where image-latent pairs that maximize the mutual information are considered as positive pairs and the rest as negative pairs. Specifically, we map the input of a generator, which was sampled from the categorical distribution, to the embedding space of the discriminator and let them act as a cluster centroid. In this way, C3-GAN succeeded in learning a clustering-friendly embedding space where each cluster is distinctively separable. Experimental results show that C3-GAN achieved the state-of-the-art clustering performance on four fine-grained image datasets, while also alleviating the mode collapse phenomenon. Code is available at https://github.com/naver-ai/c3-gan.

Learning Multimodal VAEs through Mutual Supervision

- Tom Joy, Yuge Shi, Philip Torr, Tom Rainforth, Sebastian M Schmon, Siddharth N
- abstract@open-review(Spotlight): Multimodal VAEs seek to model the joint distribution over heterogeneous data (e.g.\ vision, language), whilst also capturing a shared representation across such modalities. Prior work has typically combined information from the modalities by reconciling idiosyncratic representations directly in the recognition model through explicit products, mixtures, or other such factorisations. Here we introduce a novel alternative, the MEME, that avoids such explicit combinations by repurposing semi-supervised VAEs to combine information between modalities implicitly through mutual supervision. This formulation naturally allows learning from partially-observed data where some modalities can be entirely missing---something that most existing approaches either cannot handle, or do so to a limited extent. We demonstrate that MEME outperforms baselines on standard metrics across both partial and complete observation schemes on the MNIST-SVHN (image--image) and CUB (image--text) datasets. We also contrast the quality of the representations learnt by mutual supervision against standard approaches and observe interesting trends in its ability to capture relatedness between data.

When should agents explore?

- Miruna Pislar, David Szepesvari, Georg Ostrovski, Diana L Borsa, Tom Schaul
- abstract@open-review(Spotlight): Exploration remains a central challenge for reinforcement learning (RL). Virtually all existing methods share the feature of a *monolithic* behaviour policy that changes only gradually (at best). In contrast, the exploratory behaviours of animals and humans exhibit a rich diversity, namely including forms of *switching* between modes. This paper presents an initial study of mode-switching, non-monolithic exploration for RL. We investigate different modes to switch between, at what timescales it makes sense to switch, and what signals make for good switching triggers. We also propose practical algorithmic components that make the switching mechanism adaptive and robust, which enables flexibility without an accompanying hyper-parameter-tuning burden. Finally, we report a promising initial study on Atari, using two-mode exploration and switching at sub-episodic time-scales.

Revisiting Design Choices in Offline Model Based Reinforcement Learning

- Cong Lu, Philip Ball, Jack Parker-Holder, Michael Osborne, Stephen J. Roberts
- abstract@open-review(Spotlight): Offline reinforcement learning enables agents to leverage large pre-collected datasets of environment transitions to learn control policies, circumventing the need for potentially expensive or unsafe online data collection. Significant progress has been made recently in offline model-based reinforcement learning, approaches which leverage a learned dynamics model. This typically involves constructing a probabilistic model, and using the model uncertainty to penalize rewards where there is insufficient data, solving for a pessimistic MDP that lower bounds the true MDP. Existing methods, however, exhibit a breakdown between theory and practice, whereby pessimistic return ought to be bounded by the total variation distance of the model from the true dynamics, but is instead implemented through a penalty based on estimated model uncertainty. This has spawned a variety of uncertainty heuristics, with little to no comparison between differing approaches. In this paper, we compare these heuristics, and design novel protocols to investigate their interaction with other hyperparameters, such as the number of models, or imaginary rollout horizon. Using these insights, we show that selecting these key hyperparameters using Bayesian Optimization produces superior configurations that are vastly different to those currently used in existing hand-tuned state-of-the-art methods, and result in drastically stronger performance.

NASPY: Automated Extraction of Automated Machine Learning Models

- Xiaoxuan Lou, Shangwei Guo, Jiwei Li, Yaoxin Wu, Tianwei Zhang
- abstract@open-review(Spotlight): We present NASPY, an end-to-end adversarial framework to extract the networkarchitecture of deep learning models from Neural Architecture Search (NAS). Existing works about model extraction attacks mainly focus on conventional DNN models with very simple operations, or require heavy manual analysis with lots of domain knowledge. In contrast, NASPY introduces seq2seq models to automatically identify novel and complicated operations (e.g., separable convolution,dilated convolution) from hardware side-channel sequences. We design two models (RNN-CTC and transformer), which can achieve only 3.2% and 11.3% error rates for operation prediction. We further present methods to recover the model hyper-parameters and topology from the operation sequence . With these techniques, NASPY is able to extract the complete NAS model architecture with high fidelity and automation, which are rarely analyzed before.

COptiDICE: Offline Constrained Reinforcement Learning via Stationary Distribution Correction Estimation

- Jongmin Lee, Cosmin Paduraru, Daniel J Mankowitz, Nicolas Heess, Doina Precup, Kee-Eung Kim, Arthur Guez
- abstract@open-review(Spotlight): We consider the offline constrained reinforcement learning (RL) problem, in which the agent aims to compute a policy that maximizes expected return while satisfying given cost constraints, learning only from a pre-collected dataset. This problem setting is appealing in many real-world scenarios, where direct interaction with the environment is costly or risky, and where the resulting policy should comply with safety constraints. However, it is challenging to compute a policy that guarantees satisfying the cost constraints in the offline RL setting, since the off-policy evaluation inherently has an estimation error. In this paper, we present an offline constrained RL algorithm that optimizes the policy in the space of the stationary distribution. Our algorithm, COptiDICE, directly estimates the stationary distribution corrections of the optimal policy with respect to returns, while constraining the cost upper bound, with the goal of yielding a cost-conservative policy for actual constraint satisfaction. Experimental results show that COptiDICE attains better policies in terms of constraint satisfaction and return-maximization, outperforming baseline algorithms.

Learning Vision-Guided Quadrupedal Locomotion End-to-End with Cross-Modal Transformers

- Ruihan Yang, Minghao Zhang, Nicklas Hansen, Huazhe Xu, Xiaolong Wang
- abstract@open-review(Spotlight): We propose to address quadrupedal locomotion tasks using Reinforcement Learning (RL) with a Transformer-based model that learns to combine proprioceptive information and high-dimensional depth sensor inputs. While learning-based locomotion has made great advances using RL, most methods still rely on domain randomization for training blind agents that generalize to challenging terrains. Our key insight is that proprioceptive states only offer contact measurements for immediate reaction, whereas an agent equipped with visual sensory observations can learn to proactively maneuver environments with obstacles and uneven terrain by anticipating changes in the environment many steps ahead. In this paper, we introduce LocoTransformer, an end-to-end RL method that leverages both proprioceptive states and visual observations for locomotion control. We evaluate our method in challenging simulated environments with different obstacles and uneven terrain. We transfer our learned policy from simulation to a real robot by running it indoor and in-the-wild with unseen obstacles and terrain. Our method not only significantly improves over baselines, but also achieves far better generalization performance, especially when transferred to the real robot. Our project page with videos is at https://rchalyang.github.io/LocoTransformer/.

ViTGAN: Training GANs with Vision Transformers

- Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, Ce Liu
- abstract@open-review(Spotlight): Recently, Vision Transformers (ViTs) have shown competitive performance on image recognition while requiring less vision-specific inductive biases. In this paper, we investigate if such performance can be extended to image generation. To this end, we integrate the ViT architecture into generative adversarial networks (GANs). For ViT discriminators, we observe that existing regularization methods for GANs interact poorly with self-attention, causing serious instability during training. To resolve this issue, we introduce several novel regularization techniques for training GANs with ViTs. For ViT generators, we examine architectural choices for latent and pixel mapping layers to faciliate convergence. Empirically,

our approach, named ViTGAN, achieves comparable performance to the leading CNN- based GAN models on three datasets: CIFAR-10, CelebA, and LSUN bedroom.

POETREE: Interpretable Policy Learning with Adaptive Decision Trees

- AlizÃ©e Pace, Alex Chan, Mihaela van der Schaar
- abstract@open-review(Spotlight): Building models of human decision-making from observed behaviour is critical to better understand, diagnose and support real-world policies such as clinical care. As established policy learning approaches remain focused on imitation performance, they fall short of explaining the demonstrated decision-making process. Policy Extraction through decision Trees (POETREE) is a novel framework for interpretable policy learning, compatible with fully-offline and partially-observable clinical decision environments -- and builds probabilistic tree policies determining physician actions based on patients' observations and medical history. Fully-differentiable tree architectures are grown incrementally during optimization to adapt their complexity to the modelling task, and learn a representation of patient history through recurrence, resulting in decision tree policies that adapt over time with patient information. This policy learning method outperforms the state-of-the-art on real and synthetic medical datasets, both in terms of understanding, quantifying and evaluating observed behaviour as well as in accurately replicating it -- with potential to improve future decision support systems.

TRGP: Trust Region Gradient Projection for Continual Learning

- Sen Lin, Li Yang, Deliang Fan, Junshan Zhang
- abstract@open-review(Spotlight): Catastrophic forgetting is one of the major challenges in continual learning. To address this issue, some existing methods put restrictive constraints on the optimization space of the new task for minimizing the interference to old tasks. However, this may lead to unsatisfactory performance for the new task, especially when the new task is strongly correlated with old tasks. To tackle this challenge, we propose Trust Region Gradient Projection (TRGP) for continual learning to facilitate the forward knowledge transfer based on an efficient characterization of task correlation. Particularly, we introduce a notion of 'trust region' to select the most related old tasks for the new task in a layer-wise and single-shot manner, using the norm of gradient projection onto the subspace spanned by task inputs. Then, a scaled weight projection is proposed to cleverly reuse the frozen weights of the selected old tasks in the trust region through a layer-wise scaling matrix. By jointly optimizing the scaling matrices and the model, where the model is updated along the directions orthogonal to the subspaces of old tasks, TRGP can effectively prompt knowledge transfer without forgetting. Extensive experiments show that our approach achieves significant improvement over related state-of-the-art methods.

Properties from mechanisms: an equivariance perspective on identifiable representation learning

- Kartik Ahuja, Jason Hartford, Yoshua Bengio
- abstract@open-review(Spotlight): A key goal of unsupervised representation learning is `inverting'' a data generating process to recover its latent properties. Existing work that provably achieves this goal relies on strong assumptions on relationships between the latent variables (e.g., independence conditional on auxiliary information). In this paper, we take a very different perspective on the problem and ask,`Can we instead identify latent properties by leveraging knowledge of the mechanisms that govern their evolution?" We provide a complete characterization of the sources of non-identifiability as we vary knowledge about a set of possible mechanisms. In particular, we prove that if we know the exact mechanisms under which the latent properties evolve, then identification can be achieved up to any equivariances that are shared by the underlying mechanisms. We generalize this characterization to settings where we only know some hypothesis class over possible mechanisms, as well as settings where the mechanisms are stochastic. We demonstrate the power of this mechanism-based perspective by showing that we can leverage our results to generalize existing identifiable representation learning results. These results suggest that by exploiting inductive biases on mechanisms, it is possible to design a range of new identifiable representation learning approaches.

Revisiting Over-smoothing in BERT from the Perspective of Graph

- Han Shi, JIAHUI GAO, Hang Xu, Xiaodan Liang, Zhenguo Li, Lingpeng Kong, Stephen M. S. Lee, James Kwok
- abstract@open-review(Spotlight): Recently over-smoothing phenomenon of Transformer-based models is observed in both vision and language fields. However, no existing work has delved deeper to further investigate the main cause of this phenomenon. In this work, we make the attempt to analyze the over-smoothing problem from the perspective of graph, where such problem was first discovered and explored. Intuitively, the self-attention matrix can be seen as a normalized adjacent matrix of a corresponding graph. Based on the above connection, we provide some theoretical analysis and find that layer normalization plays a key role in the over-smoothing issue of Transformer-based models. Specifically, if the standard deviation of layer normalization is sufficiently large, the output of Transformer stacks will converge to a specific low-rank subspace and result in over-smoothing. To alleviate the over-smoothing problem, we consider hierarchical fusion strategies, which combine the representations from different layers adaptively to make the output more diverse. Extensive experiment results on various data sets illustrate the effect of our fusion method.

Training invariances and the low-rank phenomenon: beyond linear networks

- Thien Le, Stefanie Jegelka
- abstract@open-review(Spotlight): The implicit bias induced by the training of neural networks has become a topic of rigorous study. In the limit of gradient flow and gradient descent with appropriate step size, it has been shown that when one trains a deep linear network with logistic or exponential loss on linearly separable data, the weights converge to rank-1 matrices. In this paper, we extend this theoretical result to the last few linear layers of the much wider class of nonlinear ReLU-activated feedforward networks containing fully-connected layers and skip connections. Similar to the linear case, the proof relies on specific local training invariances, sometimes referred to as alignment, which we show to hold for submatrices where neurons are stably-activated in all training examples, and it reflects empirical results in the literature. We also show this is not true in general for the full matrix of ReLU fully-connected layers. Our proof relies on a specific decomposition of the network into a multilinear function and another ReLU network whose weights are constant under a certain parameter directional convergence.

Learning Long-Term Reward Redistribution via Randomized Return Decomposition

- Zhizhou Ren, Ruihan Guo, Yuan Zhou, Jian Peng
- abstract@open-review(Spotlight): Many practical applications of reinforcement learning require agents to learn from sparse and delayed rewards. It challenges the ability of agents to attribute their actions to future outcomes. In this paper, we consider the problem formulation of episodic reinforcement learning with trajectory feedback. It refers to an extreme delay of reward signals, in which the agent can only obtain one reward signal at the end of each trajectory. A popular paradigm for this problem setting is learning with a designed auxiliary dense reward function, namely proxy reward, instead of sparse environmental signals. Based on this framework, this paper proposes a novel reward redistribution algorithm, randomized return decomposition (RRD), to learn a proxy reward function for episodic reinforcement learning. We establish a surrogate problem by Monte-Carlo sampling that scales up least-squares-based reward redistribution to long-horizon problems. We analyze our surrogate loss function by connection with existing methods in the literature, which illustrates the algorithmic properties of our approach. In experiments, we extensively evaluate our proposed method on a variety of benchmark tasks with episodic rewards and demonstrate substantial improvement over baseline algorithms.

What Happens after SGD Reaches Zero Loss? --A Mathematical Framework

- Zhiyuan Li, Tianhao Wang, Sanjeev Arora
- abstract@open-review(Spotlight): Understanding the implicit bias of Stochastic Gradient Descent (SGD) is one of the key challenges in deep learning, especially for overparametrized models, where the local minimizers of the loss function L can form a manifold. Intuitively, with a sufficiently small learning rate η, SGD tracks Gradient Descent (GD) until it gets close to such manifold, where the gradient noise prevents further convergence. In such regime, Blanc et al. (2020) proved that SGD with label noise locally decreases a regularizer-like term, the sharpness of loss, $\text{tr}[\nabla^2 L]$. The current paper gives a general framework for such analysis by adapting ideas from Katzenberger (1991). It allows in principle a complete characterization for the regularization effect of SGD around such manifold---i.e., the "implicit bias"---using a stochastic differential equation (SDE) describing the limiting dynamics of the parameters, which is determined jointly by the loss function and the noise covariance. This yields some new results: (1) a *global* analysis of the implicit bias valid for η^{-2} steps, in contrast to the local analysis of Blanc et al. (2020) that is only valid for $\eta^{-1.6}$ steps and (2) allowing *arbitrary* noise covariance. As an application, we show with arbitrary large initialization, label noise SGD can always escape the kernel regime and only requires $O(\kappa\ln d)$ samples for learning an κ-sparse overparametrized linear model in $\mathbb{R}^d$ (Woodworth et al., 2020), while GD initialized in the kernel regime requires $\Omega(d)$ samples. This upper bound is minimax optimal and improves the previous $\widetilde{O}(\kappa^2)$ upper bound (HaoChen et al., 2020).

Graph-Augmented Normalizing Flows for Anomaly Detection of Multiple Time Series

- Enyan Dai, Jie Chen
- abstract@open-review(Spotlight): Anomaly detection is a widely studied task for a broad variety of data types; among them, multiple time series appear frequently in applications, including for example, power grids and traffic networks. Detecting anomalies for multiple time series, however, is a challenging subject, owing to the intricate interdependencies among the constituent series. We hypothesize that anomalies occur in low density regions of a distribution and explore the use of normalizing flows for unsupervised anomaly detection, because of their superior quality in density estimation. Moreover, we propose a novel flow model by imposing a Bayesian network among constituent series. A Bayesian network is a directed acyclic graph (DAG) that models causal relationships; it factorizes the joint probability of the series into the product of easy-to-evaluate conditional probabilities. We call such a graph-augmented normalizing flow approach GANF and propose joint estimation of the DAG with flow parameters. We conduct extensive experiments on real-world datasets and demonstrate the effectiveness of GANF for density estimation, anomaly detection, and identification of time series distribution drift.

Autoregressive Quantile Flows for Predictive Uncertainty Estimation

- Phillip Si, Allan Bishop, Volodymyr Kuleshov
- abstract@open-review(Spotlight): Numerous applications of machine learning involve representing probability distributions over high-dimensional data. We propose autoregressive quantile flows, a flexible class of normalizing flow models trained using a novel objective based on proper scoring rules. Our objective does not require calculating computationally expensive determinants of Jacobians during training and supports new types of neural architectures, such as neural autoregressive flows from which it is easy to sample. We leverage these models in quantile flow regression, an approach that parameterizes predictive conditional distributions with flows, resulting in improved probabilistic predictions on tasks such as time series forecasting and object detection. Our novel objective functions and neural flow parameterizations also yield improvements on popular generation and density estimation tasks, and represent a step beyond maximum likelihood learning of flows.

On the Importance of Firth Bias Reduction in Few-Shot Classification

- Saba Ghaffari, Ehsan Saleh, David Forsyth, Yu-Xiong Wang
- abstract@open-review(Spotlight): Learning accurate classifiers for novel categories from very few examples, known as few-shot image classification, is a challenging task in statistical machine learning and computer vision. The performance in few-shot classification suffers from the bias in the estimation of classifier parameters; however, an effective underlying bias reduction technique that could alleviate this issue in training few-shot classifiers has been overlooked. In this work, we demonstrate the effectiveness of Firth bias reduction in few-shot classification. Theoretically, Firth bias reduction removes the $O(N^{-1})$ first order term from the small-sample bias of the Maximum Likelihood Estimator. Here we show that the general Firth bias reduction technique simplifies to encouraging uniform class assignment probabilities for multinomial logistic classification, and almost has the same effect in cosine classifiers. We derive an easy-to-implement optimization objective for Firth penalized multinomial logistic and cosine classifiers, which is equivalent to penalizing the cross-entropy loss with a KL-divergence between the predictions and the uniform label distribution. Then, we empirically evaluate that it is consistently effective across the board for few-shot image classification, regardless of (1) the feature representations from different backbones, (2) the number of samples per class, and (3) the number of classes. Furthermore, we demonstrate the effectiveness of Firth bias reduction on cross-domain and imbalanced data settings. Our implementation is available at https://github.com/ehsansaleh/firth_bias_reduction.

Finite-Time Convergence and Sample Complexity of Multi-Agent Actor-Critic Reinforcement Learning with Average Reward

- FNU Hairi, Jia Liu, Songtao Lu
- abstract@open-review(Spotlight): In this paper, we establish the first finite-time convergence result of the actor-critic algorithm for fully decentralized multi-agent reinforcement learning (MARL) problems with average reward. In this problem, a set of N agents work cooperatively to maximize the global average reward through interacting with their neighbors over a communication network. We consider a practical MARL setting, where the rewards and actions of each agent are only known to itself, and the knowledge of joint actions of the agents is not assumed. Toward this end, we propose a mini-batch Markovian sampled fully decentralized actor-critic algorithm and analyze its finite-time convergence and sample complexity. We show that the sample complexity of this algorithm is $\mathcal{O}(N^{2}/\epsilon^{2}\log(N/\epsilon))$. Interestingly, this sample complexity bound matches that of the state-of-the-art single-agent actor-critic algorithms for reinforcement learning.

Towards Understanding the Data Dependency of Mixup-style Training

- Muthu Chidambaram, Xiang Wang, Yuzheng Hu, Chenwei Wu, Rong Ge
- abstract@open-review(Spotlight): In the Mixup training paradigm, a model is trained using convex combinations of data points and their associated labels. Despite seeing very few true data points during training, models trained using Mixup seem to still minimize the original empirical risk and exhibit better generalization and robustness on various tasks when compared to standard training. In this paper, we investigate how these benefits of Mixup training rely on properties of the data in the context of classification. For minimizing the original empirical risk, we compute a closed form for the Mixup-optimal classification, which allows us to construct a simple dataset on which minimizing the Mixup loss leads to learning a classifier that does not minimize the empirical loss on the data. On the other hand, we also give sufficient conditions for Mixup training to also minimize the original empirical risk. For generalization, we characterize the margin of a Mixup classifier, and use this to understand why the decision boundary of a Mixup classifier can adapt better to the full structure of the training data when compared to standard training. In contrast, we also show that, for a large class of linear models and linearly separable datasets, Mixup training leads to learning the same classifier as standard training.

Self-Supervision Enhanced Feature Selection with Correlated Gates

- Changhee Lee, Fergus Imrie, Mihaela van der Schaar
- abstract@open-review(Spotlight): Discovering relevant input features for predicting a target variable is a key scientific question. However, in many domains, such as medicine and biology, feature selection is confounded by a scarcity of labeled samples coupled with significant correlations among

features. In this paper, we propose a novel deep learning approach to feature selection that addresses both challenges simultaneously. First, we pre-train the network using unlabeled samples within a self-supervised learning framework by solving pretext tasks that require the network to learn informative representations from partial feature sets. Then, we fine-tune the pre-trained network to discover relevant features using labeled samples. During both training phases, we explicitly account for the correlation structure of the input features by generating correlated gate vectors from a multivariate Bernoulli distribution. Experiments on multiple real-world datasets including clinical and omics demonstrate that our model discovers relevant features that provide superior prediction performance compared to the state-of-the-art benchmarks in practical scenarios where there is often limited labeled data and high correlations among features.

Score-Based Generative Modeling with Critically-Damped Langevin Diffusion

- Tim Dockhorn, Arash Vahdat, Karsten Kreis
- abstract@open-review(Spotlight): Score-based generative models (SGMs) have demonstrated remarkable synthesis quality. SGMs rely on a diffusion process that gradually perturbs the data towards a tractable distribution, while the generative model learns to denoise. The complexity of this denoising task is, apart from the data distribution itself, uniquely determined by the diffusion process. We argue that current SGMs employ overly simplistic diffusions, leading to unnecessarily complex denoising processes, which limit generative modeling performance. Based on connections to statistical mechanics, we propose a novel critically-damped Langevin diffusion (CLD) and show that CLD-based SGMs achieve superior performance. CLD can be interpreted as running a joint diffusion in an extended space, where the auxiliary variables can be considered "velocities" that are coupled to the data variables as in Hamiltonian dynamics. We derive a novel score matching objective for CLD and show that the model only needs to learn the score function of the conditional distribution of the velocity given data, an easier task than learning scores of the data directly. We also derive a new sampling scheme for efficient synthesis from CLD-based diffusion models. We find that CLD outperforms previous SGMs in synthesis quality for similar network architectures and sampling compute budgets. We show that our novel sampler for CLD significantly outperforms solvers such as Eulerâ€“Maruyama. Our framework provides new insights into score-based denoising diffusion models and can be readily used for high-resolution image synthesis. Project page and code: https://nv-tlabs.github.io/CLD-SGM.

Controlling Directions Orthogonal to a Classifier

- Yilun Xu, Hao He, Tianxiao Shen, Tommi S. Jaakkola
- abstract@open-review(Spotlight): We propose to identify directions invariant to a given classifier so that these directions can be controlled in tasks such as style transfer. While orthogonal decomposition is directly identifiable when the given classifier is linear, we formally define a notion of orthogonality in the non-linear case. We also provide a surprisingly simple method for constructing the orthogonal classifier (a classifier utilizing directions other than those of the given classifier). Empirically, we present three use cases where controlling orthogonal variation is important: style transfer, domain adaptation, and fairness. The orthogonal classifier enables desired style transfer when domains vary in multiple aspects, improves domain adaptation with label shifts and mitigates the unfairness as a predictor. The code is available at https://github.com/Newbeeer/orthogonal_classifier

R5: Rule Discovery with Reinforced and Recurrent Relational Reasoning

- Shengyao Lu, Bang Liu, Keith G Mills, SHANGLING JUI, Di Niu
- abstract@open-review(Spotlight): Systematicity, i.e., the ability to recombine known parts and rules to form new sequences while reasoning over relational data, is critical to machine intelligence. A model with strong systematicity is able to train on small-scale tasks and generalize to large-scale tasks. In this paper, we propose R5, a relational reasoning framework based on reinforcement learning that reasons over relational graph data and explicitly mines underlying compositional logical rules from observations. R5 has strong systematicity and being robust to noisy data. It consists of a policy value network equipped with Monte Carlo Tree Search to perform recurrent relational prediction and a backtrack rewriting mechanism for rule mining. By alternately applying the two components, R5 progressively learns a set of explicit rules from data and performs explainable and generalizable relation prediction. We conduct extensive evaluations on multiple datasets. Experimental results show that R5 outperforms various embedding-based and rule induction baselines on relation prediction tasks while achieving a high recall rate in discovering ground truth rules.

Representation Learning for Online and Offline RL in Low-rank MDPs

- Masatoshi Uehara, Xuezhou Zhang, Wen Sun
- abstract@open-review(Spotlight): This work studies the question of Representation Learning in RL: how can we learn a compact low-dimensional representation such that on top of the representation we can perform RL procedures such as exploration and exploitation, in a sample efficient manner. We focus on the low-rank Markov Decision Processes (MDPs) where the transition dynamics correspond to a low-rank transition matrix. Unlike prior works that assume the representation is known (e.g., linear MDPs), here we need to learn the representation for the low-rank MDP. We study both the online RL and offline RL settings. For the online setting, operating with the same computational oracles used in FLAMBE (Agarwal et.al), the state-of-art algorithm for learning representations in low-rank MDPs, we propose an algorithm REP-UCB Upper Confidence Bound driven Representation learning for RL), which significantly improves the sample complexity from $\widetilde{O}(A^9 d^7 / (\epsilon^{10} (1-\gamma)^{22}))$ for FLAMBE to $\widetilde{O}(A^4 d^4 / (\epsilon^2 (1-\gamma)^{2}))$ with d being the rank of the transition matrix (or dimension of the ground truth representation), A being the number of actions, and γ being the discounted factor. Notably, REP-UCB is simpler than FLAMBE, as it directly balances the interplay between representation learning, exploration, and exploitation, while FLAMBE is an explore-then-commit style approach and has to perform reward-free exploration step-by-step forward in time. For the offline RL setting, we develop an algorithm that leverages pessimism to learn under a partial coverage condition: our algorithm is able to compete against any policy as long as it is covered by the offline distribution.

Lossless Compression with Probabilistic Circuits

- Anji Liu, Stephan Mandt, Guy Van den Broeck
- abstract@open-review(Spotlight): Despite extensive progress on image generation, common deep generative model architectures are not easily applied to lossless compression. For example, VAEs suffer from a compression cost overhead due to their latent variables. This overhead can only be partially eliminated with elaborate schemes such as bits-back coding, often resulting in poor single-sample compression rates. To overcome such problems, we establish a new class of tractable lossless compression models that permit efficient encoding and decoding: Probabilistic Circuits (PCs). These are a class of neural networks involving $|p|$ computational units that support efficient marginalization over arbitrary subsets of the D feature dimensions, enabling efficient arithmetic coding. We derive efficient encoding and decoding schemes that both have time complexity $\mathcal{O} (\log(D) \cdot |p|)$, where a naive scheme would have linear costs in D and $|p|$, making the approach highly scalable. Empirically, our PC-based (de)compression algorithm runs 5-40 times faster than neural compression algorithms that achieve similar bitrates. By scaling up the traditional PC structure learning pipeline, we achieve state-of-the-art results on image datasets such as MNIST. Furthermore, PCs can be naturally integrated with existing neural compression algorithms to improve the performance of these base models on natural image datasets. Our results highlight the potential impact that non-standard learning architectures may have on neural data compression.

Understanding Domain Randomization for Sim-to-real Transfer

- Xiaoyu Chen, Jiachen Hu, Chi Jin, Lihong Li, Liwei Wang
- abstract@open-review(Spotlight): Reinforcement learning encounters many challenges when applied directly in the real world. Sim-to-real transfer is widely used to transfer the knowledge learned from simulation to the real world. Domain randomization---one of the most popular algorithms for sim-to-

real transfer---has been demonstrated to be effective in various tasks in robotics and autonomous driving. Despite its empirical successes, theoretical understanding on why this simple algorithm works is largely missing. In this paper, we propose a theoretical framework for sim-to-real transfers, in which the simulator is modeled as a set of MDPs with tunable parameters (corresponding to unknown physical parameters such as friction). We provide sharp bounds on the sim-to-real gap---the difference between the value of policy returned by domain randomization and the value of an optimal policy for the real world. We prove that sim-to-real transfer can succeed under mild conditions without any real-world training samples. Our theory also highlights the importance of using memory (i.e., history-dependent policies) in domain randomization. Our proof is based on novel techniques that reduce the problem of bounding the sim-to-real gap to the problem of designing efficient learning algorithms for infinite-horizon MDPs, which we believe are of independent interest.

$\mathrm{SO}(2)$-Equivariant Reinforcement Learning

- Dian Wang, Robin Walters, Robert Platt
- abstract@open-review(Spotlight): Equivariant neural networks enforce symmetry within the structure of their convolutional layers, resulting in a substantial improvement in sample efficiency when learning an equivariant or invariant function. Such models are applicable to robotic manipulation learning which can often be formulated as a rotationally symmetric problem. This paper studies equivariant model architectures in the context of Q-learning and actor-critic reinforcement learning. We identify equivariant and invariant characteristics of the optimal Q-function and the optimal policy and propose equivariant DQN and SAC algorithms that leverage this structure. We present experiments that demonstrate that our equivariant versions of DQN and SAC can be significantly more sample efficient than competing algorithms on an important class of robotic manipulation problems.

Scarf: Self-Supervised Contrastive Learning using Random Feature Corruption

- Dara Bahri, Heinrich Jiang, Yi Tay, Donald Metzler
- abstract@open-review(Spotlight): Self-supervised contrastive representation learning has proved incredibly successful in the vision and natural language domains, enabling state-of-the-art performance with orders of magnitude less labeled data. However, such methods are domain-specific and little has been done to leverage this technique on real-world \emph{tabular} datasets. We propose \textsc{Scarf}, a simple, widely-applicable technique for contrastive learning, where views are formed by corrupting a random subset of features. When applied to pre-train deep neural networks on the 69 real-world, tabular classification datasets from the OpenML-CC18 benchmark, \textsc{Scarf} not only improves classification accuracy in the fully-supervised setting but does so also in the presence of label noise and in the semi-supervised setting where only a fraction of the available training data is labeled. We show that \textsc{Scarf} complements existing strategies and outperforms alternatives like autoencoders. We conduct comprehensive ablations, detailing the importance of a range of factors.

Responsible Disclosure of Generative Models Using Scalable Fingerprinting

- Ning Yu, Vladislav Skripniuk, Dingfan Chen, Larry S. Davis, Mario Fritz
- abstract@open-review(Spotlight): Over the past years, deep generative models have achieved a new level of performance. Generated data has become difficult, if not impossible, to be distinguished from real data. While there are plenty of use cases that benefit from this technology, there are also strong concerns on how this new technology can be misused to generate deep fakes and enable misinformation at scale. Unfortunately, current deep fake detection methods are not sustainable, as the gap between real and fake continues to close. In contrast, our work enables a responsible disclosure of such state-of-the-art generative models, that allows model inventors to fingerprint their models, so that the generated samples containing a fingerprint can be accurately detected and attributed to a source. Our technique achieves this by an efficient and scalable ad-hoc generation of a large population of models with distinct fingerprints. Our recommended operation point uses a 128-bit fingerprint which in principle results in more than 10^{38} identifiable models. Experiments show that our method fulfills key properties of a fingerprinting mechanism and achieves effectiveness in deep fake detection and attribution. Code and models are available at https://github.com/ningyu1991/ScalableGANFingerprints.

Path Auxiliary Proposal for MCMC in Discrete Space

- Haoran Sun, Hanjun Dai, Wei Xia, Arun Ramamurthy
- abstract@open-review(Spotlight): Energy-based Model (EBM) offers a powerful approach for modeling discrete structure, but both inference and learning of EBM are hard as it involves sampling from discrete distributions. Recent work shows Markov Chain Monte Carlo (MCMC) with the informed proposal is a powerful tool for such sampling. However, an informed proposal only allows local updates as it requires evaluating all energy changes in the neighborhood. In this work, we present a path auxiliary algorithm that uses a composition of local moves to efficiently explore large neighborhoods. We also give a fast version of our algorithm that only queries the evaluation of energy function twice for each proposal via linearization of the energy function. Empirically, we show that our path auxiliary algorithms considerably outperform other generic samplers on various discrete models for sampling, inference, and learning. Our method can also be used to train deep EBMs for high-dimensional discrete data.

Possibility Before Utility: Learning And Using Hierarchical Affordances

- Robby Costales, Shariq Iqbal, Fei Sha
- abstract@open-review(Spotlight): Reinforcement learning algorithms struggle on tasks with complex hierarchical dependency structures. Humans and other intelligent agents do not waste time assessing the utility of every high-level action in existence, but instead only consider ones they deem possible in the first place. By focusing only on what is feasible, or "afforded", at the present moment, an agent can spend more time both evaluating the utility of and acting on what matters. To this end, we present Hierarchical Affordance Learning (HAL), a method that learns a model of hierarchical affordances in order to prune impossible subtasks for more effective learning. Existing works in hierarchical reinforcement learning provide agents with structural representations of subtasks but are not affordance-aware, and by grounding our definition of hierarchical affordances in the present state, our approach is more flexible than the multitude of approaches that ground their subtask dependencies in a symbolic history. While these logic-based methods often require complete knowledge of the subtask hierarchy, our approach is able to utilize incomplete and varying symbolic specifications. Furthermore, we demonstrate that relative to non-affordance-aware methods, HAL agents are better able to efficiently learn complex tasks, navigate environment stochasticity, and acquire diverse skills in the absence of extrinsic supervision---all of which are hallmarks of human learning.

Interpretable Unsupervised Diversity Denoising and Artefact Removal

- Mangal Prakash, Mauricio Delbracio, Peyman Milanfar, Florian Jug
- abstract@open-review(Spotlight): Image denoising and artefact removal are complex inverse problems admitting multiple valid solutions. Unsupervised diversity restoration, that is, obtaining a diverse set of possible restorations given a corrupted image, is important for ambiguity removal in many applications such as microscopy where paired data for supervised training are often unobtainable. In real world applications, imaging noise and artefacts are typically hard to model, leading to unsatisfactory performance of existing unsupervised approaches. This work presents an interpretable approach for unsupervised and diverse image restoration. To this end, we introduce a capable architecture called Hierarchical DivNoising (HDN) based on hierarchical Variational Autoencoder. We show that HDN learns an interpretable multi-scale representation of artefacts and we leverage this interpretability to remove imaging artefacts commonly occurring in microscopy data. Our method achieves state-of-the-art results on twelve benchmark image denoising datasets while providing access to a whole distribution of sensibly restored solutions. Additionally, we demonstrate on three real microscopy datasets that HDN removes artefacts without supervision, being the first method capable of doing so while generating multiple plausible restorations all consistent with the given corrupted image.

Half-Inverse Gradients for Physical Deep Learning

- Patrick Schnell, Philipp Holl, Nils Thuerey
- abstract@open-review(Spotlight): Recent works in deep learning have shown that integrating differentiable physics simulators into the training process can greatly improve the quality of results. Although this combination represents a more complex optimization task than usual neural network training, the same gradient-based optimizers are used to minimize the loss function. However, the integrated physics solvers have a profound effect on the gradient flow as manipulating scales in magnitude and direction is an inherent property of many physical processes. Consequently, the gradient flow is often highly unbalanced and creates an environment in which existing gradient-based optimizers perform poorly. In this work, we analyze the characteristics of both physical and neural network optimizations separately to derive a new method based on a half-inversion of the Jacobian. Our approach combines principles of both classical network and physics optimizers to solve the combined optimization task. Compared to state-of-the-art neural network optimizers, our method converges more quickly and to better solutions, which we demonstrate on three complex learning problems involving nonlinear oscillators, the Schroedinger equation and the Poisson problem.

EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits

- Yikun Ban, Yuchen Yan, Arindam Banerjee, Jingrui He
- abstract@open-review(Spotlight): In this paper, we propose a novel neural exploration strategy in contextual bandits, EE-Net, distinct from the standard UCB-based and TS-based approaches. Contextual multi-armed bandits have been studied for decades with various applications. To solve the exploitation-exploration tradeoff in bandits, there are three main techniques: epsilon-greedy, Thompson Sampling (TS), and Upper Confidence Bound (UCB). In recent literature, linear contextual bandits have adopted ridge regression to estimate the reward function and combine it with TS or UCB strategies for exploration. However, this line of works explicitly assumes the reward is based on a linear function of arm vectors, which may not be true in real-world datasets. To overcome this challenge, a series of neural bandit algorithms have been proposed, where a neural network is used to learn the underlying reward function and TS or UCB are adapted for exploration. Instead of calculating a large-deviation based statistical bound for exploration like previous methods, we propose "EE-Net", a novel neural-based exploration strategy. In addition to using a neural network (Exploitation network) to learn the reward function, EE-Net uses another neural network (Exploration network) to adaptively learn potential gains compared to the currently estimated reward for exploration. Then, a decision-maker is constructed to combine the outputs from the Exploitation and Exploration networks. We prove that EE-Net can achieve $\mathcal{O}(\sqrt{T\log T})$ regret and show that EE-Net outperforms existing linear and neural contextual bandit baselines on real-world datasets.

Spike-inspired rank coding for fast and accurate recurrent neural networks

- Alan Jeffares, Qinghai Guo, Pontus Stenetorp, Timoleon Moraitis
- abstract@open-review(Spotlight): Biological spiking neural networks (SNNs) can temporally encode information in their outputs, e.g. in the rank order in which neurons fire, whereas artificial neural networks (ANNs) conventionally do not. As a result, models of SNNs for neuromorphic computing are regarded as potentially more rapid and efficient than ANNs when dealing with temporal input. On the other hand, ANNs are simpler to train, and usually achieve superior performance. Here we show that temporal coding such as rank coding (RC) inspired by SNNs can also be applied to conventional ANNs such as LSTMs, and leads to computational savings and speedups. In our RC for ANNs, we apply backpropagation through time using the standard real-valued activations, but only from a strategically early time step of each sequential input example, decided by a threshold-crossing event. Learning then incorporates naturally also when to produce an output, without other changes to the model or the algorithm. Both the forward and the backward training pass can be significantly shortened by skipping the remaining input sequence after that first event. RC-training also significantly reduces time-to-insight during inference, with a minimal decrease in accuracy. The desired speed-accuracy trade-off is tunable by varying the threshold or a regularization parameter that rewards output entropy. We demonstrate these in two toy problems of sequence classification, and in a temporally-encoded MNIST dataset where our RC model achieves 99.19% accuracy after the first input time-step, outperforming the state of the art in temporal coding with SNNs, as well as in spoken-word classification of Google Speech Commands, outperforming non-RC-trained early inference with LSTMs.

How to Robustify Black-Box ML Models? A Zeroth-Order Optimization Perspective

- Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jinfeng Yi, Mingyi Hong, Shiyu Chang, Sijia Liu
- abstract@open-review(Spotlight): The lack of adversarial robustness has been recognized as an important issue for state-of-the-art machine learning (ML) models, e.g., deep neural networks (DNNs). Thereby, robustifying ML models against adversarial attacks is now a major focus of research. However, nearly all existing defense methods, particularly for robust training, made the white-box assumption that the defender has the access to the details of an ML model (or its surrogate alternatives if available), e.g., its architectures and parameters. Beyond existing works, in this paper we aim to address the problem of black-box defense: How to robustify a black-box model using just input queries and output feedback? Such a problem arises in practical scenarios, where the owner of the predictive model is reluctant to share model information in order to preserve privacy. To this end, we propose a general notion of defensive operation that can be applied to black-box models, and design it through the lens of denoised smoothing (DS), a ï¬ rst-order (FO) certiï¬ ed defense technique. To allow the design of merely using model queries, we further integrate DS with the zeroth-order (gradient-free) optimization. However, a direct implementation of zeroth-order (ZO) optimization suffers a high variance of gradient estimates, and thus leads to ineffective defense. To tackle this problem, we next propose to prepend an autoencoder (AE) to a given (black-box) model so that DS can be trained using variance-reduced ZO optimization. We term the eventual defense as ZO-AE-DS. In practice, we empirically show that ZO-AE-DS can achieve improved accuracy, certiï¬ ed robustness, and query complexity over existing baselines. And the effectiveness of our approach is justiï¬ ed under both image classiï¬ cation and image reconstruction tasks.

RelaxLoss: Defending Membership Inference Attacks without Losing Utility

- Dingfan Chen, Ning Yu, Mario Fritz
- abstract@open-review(Spotlight): As a long-term threat to the privacy of training data, membership inference attacks (MIAs) emerge ubiquitously in machine learning models. Existing works evidence strong connection between the distinguishability of the training and testing loss distributions and the model's vulnerability to MIAs. Motivated by existing results, we propose a novel training framework based on a relaxed loss ($\textbf{RelaxLoss}$) with a more achievable learning target, which leads to narrowed generalization gap and reduced privacy leakage. RelaxLoss is applicable to any classification model with added benefits of easy implementation and negligible overhead. Through extensive evaluations on five datasets with diverse modalities (images, medical data, transaction records), our approach consistently outperforms state-of-the-art defense mechanisms in terms of resilience against MIAs as well as model utility. Our defense is the first that can withstand a wide range of attacks while preserving (or even improving) the target model's utility.

Analyzing and Improving the Optimization Landscape of Noise-Contrastive Estimation

- Bingbin Liu, Elan Rosenfeld, Pradeep Kumar Ravikumar, Andrej Risteski
- abstract@open-review(Spotlight): Noise-contrastive estimation (NCE) is a statistically consistent method for learning unnormalized probabilistic models. It has been empirically observed that the choice of the noise distribution is crucial for NCEâ€™s performance. However, such observation has never been made formal or quantitative. In fact, it is not even clear whether the difficulties arising from a poorly chosen noise distribution are statistical or algorithmic in nature. In this work, we formally pinpoint reasons for NCEâ€™s poor performance when an inappropriate noise distribution is used. Namely, we prove

these challenges arise due to an ill-behaved (more precisely, flat) loss landscape. To address this, we introduce a variant of NCE called \emph{eNCE} which uses an exponential loss and for which \emph{normalized gradient descent} addresses the landscape issues \emph{provably} when the target and noise distributions are in a given exponential family.

Contact Points Discovery for Soft-Body Manipulations with Differentiable Physics

- Sizhe Li, Zhiao Huang, Tao Du, Hao Su, Joshua B. Tenenbaum, Chuang Gan
- abstract@open-review(Spotlight): Differentiable physics has recently been shown as a powerful tool for solving soft-body manipulation tasks. However, the differentiable physics solver often gets stuck when the initial contact points of the end effectors are sub-optimal or when performing multi-stage tasks that require contact point switching, which often leads to local minima. To address this challenge, we propose a contact point discovery approach (CPDeform) that guides the stand-alone differentiable physics solver to deform various soft-body plasticines. The key idea of our approach is to integrate optimal transport-based contact points discovery into the differentiable physics solver to overcome the local minima from initial contact points or contact switching. On single-stage tasks, our method can automatically find suitable initial contact points based on transport priorities. On complex multi-stage tasks, we can iteratively switch the contact points of end-effectors based on transport priorities. To evaluate the effectiveness of our method, we introduce PlasticineLab-M that extends the existing differentiable physics benchmark PlasticineLab to seven new challenging multi-stage soft-body manipulation tasks. Extensive experimental results suggest that: 1) on multi-stage tasks that are infeasible for the vanilla differentiable physics solver, our approach discovers contact points that efficiently guide the solver to completion; 2) on tasks where the vanilla solver performs sub-optimally or near-optimally, our contact point discovery method performs better than or on par with the manipulation performance obtained with handcrafted contact points.

Leveraging Automated Unit Tests for Unsupervised Code Translation

- Baptiste Roziere, Jie Zhang, Francois Charton, Mark Harman, Gabriel Synnaeve, Guillaume Lample
- abstract@open-review(Spotlight): With little to no parallel data available for programming languages, unsupervised methods are well-suited to source code translation. However, the majority of unsupervised machine translation approaches rely on back-translation, a method developed in the context of natural language translation and one that inherently involves training on noisy inputs. Unfortunately, source code is highly sensitive to small changes; a single token can result in compilation failures or erroneous programs, unlike natural languages where small inaccuracies may not change the meaning of a sentence. To address this issue, we propose to leverage an automated unit-testing system to filter out invalid translations, thereby creating a fully tested parallel corpus. We found that fine-tuning an unsupervised model with this filtered data set significantly reduces the noise in the translations so-generated, comfortably outperforming the state-of-the-art for all language pairs studied. In particular, for Javaâ†'Python and Pythonâ†'C++ we outperform the best previous methods by more than 16% and 24% respectively, reducing the error rate by more than 35%.

Scalable Sampling for Nonsymmetric Determinantal Point Processes

- Insu Han, Mike Gartrell, Jennifer Gillenwater, Elvis Dohmatob, amin karbasi
- abstract@open-review(Spotlight): A determinantal point process (DPP) on a collection of M items is a model, parameterized by a symmetric kernel matrix, that assigns a probability to every subset of those items. Recent work shows that removing the kernel symmetry constraint, yielding nonsymmetric DPPs (NDPPs), can lead to significant predictive performance gains for machine learning applications. However, existing work leaves open the question of scalable NDPP sampling. There is only one known DPP sampling algorithm, based on Cholesky decomposition, that can directly apply to NDPPs as well. Unfortunately, its runtime is cubic in M, and thus does not scale to large item collections. In this work, we first note that this algorithm can be transformed into a linear-time one for kernels with low-rank structure. Furthermore, we develop a scalable sublinear-time rejection sampling algorithm by constructing a novel proposal distribution. Additionally, we show that imposing certain structural constraints on the NDPP kernel enables us to bound the rejection rate in a way that depends only on the kernel rank. In our experiments we compare the speed of all of these samplers for a variety of real-world tasks.

Amortized Tree Generation for Bottom-up Synthesis Planning and Synthesizable Molecular Design

- Wenhao Gao, RocÃo Mercado, Connor W. Coley
- abstract@open-review(Spotlight): Molecular design and synthesis planning are two critical steps in the process of molecular discovery that we propose to formulate as a single shared task of conditional synthetic pathway generation. We report an amortized approach to generate synthetic pathways as a Markov decision process conditioned on a target molecular embedding. This approach allows us to conduct synthesis planning in a bottom-up manner and design synthesizable molecules by decoding from optimized conditional codes, demonstrating the potential to solve both problems of design and synthesis simultaneously. The approach leverages neural networks to probabilistically model the synthetic trees, one reaction step at a time, according to reactivity rules encoded in a discrete action space of reaction templates. We train these networks on hundreds of thousands of artificial pathways generated from a pool of purchasable compounds and a list of expert-curated templates. We validate our method with (a) the recovery of molecules using conditional generation, (b) the identification of synthesizable structural analogs, and (c) the optimization of molecular structures given oracle functions relevant to bioactivity and drug discovery.

Iterative Refinement Graph Neural Network for Antibody Sequence-Structure Co-design

- Wengong Jin, Jeremy Wohlwend, Regina Barzilay, Tommi S. Jaakkola
- abstract@open-review(Spotlight): Antibodies are versatile proteins that bind to pathogens like viruses and stimulate the adaptive immune system. The specificity of antibody binding is determined by complementarity-determining regions (CDRs) at the tips of these Y-shaped proteins. In this paper, we propose a generative model to automatically design the CDRs of antibodies with enhanced binding specificity or neutralization capabilities. Previous generative approaches formulate protein design as a structure-conditioned sequence generation task, assuming the desired 3D structure is given a priori. In contrast, we propose to co-design the sequence and 3D structure of CDRs as graphs. Our model unravels a sequence autoregressively while iteratively refining its predicted global structure. The inferred structure in turn guides subsequent residue choices. For efficiency, we model the conditional dependence between residues inside and outside of a CDR in a coarse-grained manner. Our method achieves superior log-likelihood on the test set and outperforms previous baselines in designing antibodies capable of neutralizing the SARS-CoV-2 virus.

Churn Reduction via Distillation

- Heinrich Jiang, Harikrishna Narasimhan, Dara Bahri, Andrew Cotter, Afshin Rostamizadeh
- abstract@open-review(Spotlight): In real-world systems, models are frequently updated as more data becomes available, and in addition to achieving high accuracy, the goal is to also maintain a low difference in predictions compared to the base model (i.e. predictive churn). If model retraining results in vastly different behavior, then it could cause negative effects in downstream systems, especially if this churn can be avoided with limited impact on model accuracy. In this paper, we show an equivalence between training with distillation using the base model as the teacher and training with an explicit constraint on the predictive churn. We then show that distillation performs strongly for low churn training against a number of recent baselines on a wide range of datasets and model architectures, including fully-connected networks, convolutional networks, and transformers.

Learning Causal Models from Conditional Moment Restrictions by Importance Weighting

- Masahiro Kato, Masaaki Imaizumi, Kenichiro McAlinn, Shota Yasui, Haruo Kakehi
- abstract@open-review(Spotlight): We consider learning causal relationships under conditional moment restrictions. Unlike causal inference under unconditional moment restrictions, conditional moment restrictions pose serious challenges for causal inference. To address this issue, we propose a method that transforms conditional moment restrictions to unconditional moment restrictions through importance weighting using a conditional density ratio estimator. Then, using this transformation, we propose a method that successfully estimate a parametric or nonparametric functions defined under the conditional moment restrictions. We analyze the estimation error and provide a bound on the structural function, providing theoretical support for our proposed method. In experiments, we confirm the soundness of our proposed method.

Scalable One-Pass Optimisation of High-Dimensional Weight-Update Hyperparameters by Implicit Differentiation

- Ross M Clarke, Elre Talea Oldewage, JosÃ© Miguel HernÃ¡ndez-Lobato
- abstract@open-review(Spotlight): Machine learning training methods depend plentifully and intricately on hyperparameters, motivating automated strategies for their optimisation. Many existing algorithms restart training for each new hyperparameter choice, at considerable computational cost. Some hypergradient-based one-pass methods exist, but these either cannot be applied to arbitrary optimiser hyperparameters (such as learning rates and momenta) or take several times longer to train than their base models. We extend these existing methods to develop an approximate hypergradient-based hyperparameter optimiser which is applicable to any continuous hyperparameter appearing in a differentiable model weight update, yet requires only one training episode, with no restarts. We also provide a motivating argument for convergence to the true hypergradient, and perform tractable gradient-based optimisation of independent learning rates for each model parameter. Our method performs competitively from varied random hyperparameter initialisations on several UCI datasets and Fashion-MNIST (using a one-layer MLP), Penn Treebank (using an LSTM) and CIFAR-10 (using a ResNet-18), in time only 2-3x greater than vanilla training.

Sample Efficient Deep Reinforcement Learning via Uncertainty Estimation

- Vincent Mai, Kaustubh Mani, Liam Paull
- abstract@open-review(Spotlight): In model-free deep reinforcement learning (RL) algorithms, using noisy value estimates to supervise policy evaluation and optimization is detrimental to the sample efficiency. As this noise is heteroscedastic, its effects can be mitigated using uncertainty-based weights in the optimization process. Previous methods rely on sampled ensembles, which do not capture all aspects of uncertainty. We provide a systematic analysis of the sources of uncertainty in the noisy supervision that occurs in RL, and introduce inverse-variance RL, a Bayesian framework which combines probabilistic ensembles and Batch Inverse Variance weighting. We propose a method whereby two complementary uncertainty estimation methods account for both the Q-value and the environment stochasticity to better mitigate the negative impacts of noisy supervision. Our results show significant improvement in terms of sample efficiency on discrete and continuous control tasks.

Learning Pruning-Friendly Networks via Frank-Wolfe: One-Shot, Any-Sparsity, And No Retraining

- Lu Miao, Xiaolong Luo, Tianlong Chen, Wuyang Chen, Dong Liu, Zhangyang Wang
- abstract@open-review(Spotlight): We present a novel framework to train a large deep neural network (DNN) for only $\textit{once}$, which can then be pruned to $\textit{any sparsity ratio}$ to preserve competitive accuracy $\textit{without any re-training}$. Conventional methods often require (iterative) pruning followed by re-training, which not only incurs large overhead beyond the original DNN training but also can be sensitive to retraining hyperparameters. Our core idea is to re-cast the DNN training as an explicit $\textit{pruning-aware}$ process: that is formulated with an auxiliary K-sparse polytope constraint, to encourage network weights to lie in a convex hull spanned by K-sparse vectors, potentially resulting in more sparse weight matrices. We then leverage a stochastic Frank-Wolfe (SFW) algorithm to solve this new constrained optimization, which naturally leads to sparse weight updates each time. We further note an overlooked fact that existing DNN initializations were derived to enhance SGD training (e.g., avoid gradient explosion or collapse), but was unaligned with the challenges of training with SFW. We hence also present the first learning-based initialization scheme specifically for boosting SFW-based DNN training. Experiments on CIFAR-10 and Tiny-ImageNet datasets demonstrate that our new framework named $\textbf{SFW-pruning}$ consistently achieves the state-of-the-art performance on various benchmark DNNs over a wide range of pruning ratios. Moreover, SFW-pruning only needs to train once on the same model and dataset, for obtaining arbitrary ratios, while requiring neither iterative pruning nor retraining. All codes will be released to the public.

Learning transferable motor skills with hierarchical latent mixture policies

- Dushyant Rao, Fereshteh Sadeghi, Leonard Hasenclever, Markus Wulfmeier, Martina Zambelli, Giulia Vezzani, Dhruva Tirumala, Yusuf Aytar, Josh Merel, Nicolas Heess, raia hadsell
- abstract@open-review(Spotlight): For robots operating in the real world, it is desirable to learn reusable abstract behaviours that can effectively be transferred across numerous tasks and scenarios. We propose an approach to learn skills from data using a hierarchical mixture latent variable model. Our method exploits a multi-level hierarchy of both discrete and continuous latent variables, to model a discrete set of abstract high-level behaviours while allowing for variance in how they are executed. We demonstrate in manipulation domains that the method can effectively cluster offline data into distinct, executable behaviours, while retaining the flexibility of a continuous latent variable model. The resulting skills can be transferred to new tasks, unseen objects, and from state to vision-based policies, yielding significantly better sample efficiency and asymptotic performance compared to existing skill- and imitation-based methods. We also perform further analysis showing how and when the skills are most beneficial: they encourage directed exploration to cover large regions of the state space relevant to the task, making them most effective in challenging sparse-reward settings.

Compositional Training for End-to-End Deep AUC Maximization

- Zhuoning Yuan, Zhishuai Guo, Nitesh Chawla, Tianbao Yang
- abstract@open-review(Spotlight): Recently, deep AUC maximization (DAM) has achieved great success in different domains (e.g., medical image classification). However, the end-to-end training for deep AUC maximization still remains a challenging problem. Previous studies employ an ad-hoc two-stage approach that first trains the network by optimizing a traditional loss (e.g., cross-entropy loss) and then finetunes the network by optimizing an AUC loss. This is because that training a deep neural network from scratch by maximizing an AUC loss usually does not yield a satisfactory performance. This phenomenon can be attributed to the degraded feature representations learned by maximizing the AUC loss from scratch. To address this issue, we propose a novel compositional training framework for end-to-end DAM, namely compositional DAM. The key idea of compositional training is to minimize a compositional objective function, where the outer function corresponds to an AUC loss and the inner function represents a gradient descent step for minimizing a traditional loss, e.g., the cross-entropy (CE) loss. To optimize the non-standard compositional objective, we propose an efficient and provable stochastic optimization algorithm. The proposed algorithm enhances the capabilities of both robust feature learning and robust classifier learning by alternatively taking a gradient descent step for the CE loss and for the AUC loss in a systematic way. We conduct extensive empirical studies on imbalanced benchmark and medical image datasets, which unanimously verify the effectiveness of the proposed method. Our results show that the compositional training approach dramatically improves both the feature representations and the testing AUC score compared with traditional deep learning approaches, and yields better performance than the two-stage approaches for DAM as well. The proposed method is implemented in our open-sourced library LibAUC (https://www.libauc.org) and code is available at https://github.com/Optimization-AI/LibAUC.

Explanations of Black-Box Models based on Directional Feature Interactions

- Aria Masoomi, Davin Hill, Zhonghui Xu, Craig P Hersh, Edwin K. Silverman, Peter J. Castaldi, Stratis Ioannidis, Jennifer Dy

- abstract@open-review(Spotlight): As machine learning algorithms are deployed ubiquitously to a variety of domains, it is imperative to make these often black-box models transparent. Several recent works explain black-box models by capturing the most influential features for prediction per instance; such explanation methods are univariate, as they characterize importance per feature. We extend univariate explanation to a higher-order; this enhances explainability, as bivariate methods can capture feature interactions in black-box models, represented as a directed graph. Analyzing this graph enables us to discover groups of features that are equally important (i.e., interchangeable), while the notion of directionality allows us to identify the most influential features. We apply our bivariate method on Shapley value explanations, and experimentally demonstrate the ability of directional explanations to discover feature interactions. We show the superiority of our method against state-of-the-art on CIFAR10, IMDB, Census, Divorce, Drug, and gene data.

On Lottery Tickets and Minimal Task Representations in Deep Reinforcement Learning

- Marc Vischer, Robert Tjarko Lange, Henning Sprekeler
- abstract@open-review(Spotlight): The lottery ticket hypothesis questions the role of overparameterization in supervised deep learning. But how is the performance of winning lottery tickets affected by the distributional shift inherent to reinforcement learning problems? In this work, we address this question by comparing sparse agents who have to address the non-stationarity of the exploration-exploitation problem with supervised agents trained to imitate an expert. We show that feed-forward networks trained with behavioural cloning compared to reinforcement learning can be pruned to higher levels of sparsity without performance degradation. This suggests that in order to solve the RL-specific distributional shift agents require more degrees of freedom. Using a set of carefully designed baseline conditions, we find that the majority of the lottery ticket effect in both learning paradigms can be attributed to the identified mask rather than the weight initialization. The input layer mask selectively prunes entire input dimensions that turn out to be irrelevant for the task at hand. At a moderate level of sparsity the mask identified by iterative magnitude pruning yields minimal task-relevant representations, i.e., an interpretable inductive bias. Finally, we propose a simple initialization rescaling which promotes the robust identification of sparse task representations in low-dimensional control tasks.

Self-supervised Learning is More Robust to Dataset Imbalance

- Hong Liu, Jeff Z. HaoChen, Adrien Gaidon, Tengyu Ma
- abstract@open-review(Spotlight): Self-supervised learning (SSL) is a scalable way to learn general visual representations since it learns without labels. However, large-scale unlabeled datasets in the wild often have long-tailed label distributions, where we know little about the behavior of SSL. In this work, we systematically investigate self-supervised learning under dataset imbalance. First, we find via extensive experiments that off-the-shelf self-supervised representations are already more robust to class imbalance than supervised representations. The performance gap between balanced and imbalanced pre-training with SSL is significantly smaller than the gap with supervised learning, across sample sizes, for both in-domain and, especially, out-of-domain evaluation. Second, towards understanding the robustness of SSL, we hypothesize that SSL learns richer features from frequent data: it may learn label-irrelevant-but-transferable features that help classify the rare classes and downstream tasks. In contrast, supervised learning has no incentive to learn features irrelevant to the labels from frequent examples. We validate this hypothesis with semi-synthetic experiments as well as rigorous mathematical analyses on a simplified setting. Third, inspired by the theoretical insights, we devise a re-weighted regularization technique that consistently improves the SSL representation quality on imbalanced datasets with several evaluation criteria, closing the small gap between balanced and imbalanced datasets with the same number of examples.

Ab-Initio Potential Energy Surfaces by Pairing GNNs with Neural Wave Functions

- Nicholas Gao, Stephan GÃ¼nnemann
- abstract@open-review(Spotlight): Solving the SchrÃ¶dinger equation is key to many quantum mechanical properties. However, an analytical solution is only tractable for single-electron systems. Recently, neural networks succeeded at modelling wave functions of many-electron systems. Together with the variational Monte-Carlo (VMC) framework, this led to solutions on par with the best known classical methods. Still, these neural methods require tremendous amounts of computational resources as one has to train a separate model for each molecular geometry. In this work, we combine a Graph Neural Network (GNN) with a neural wave function to simultaneously solve the SchrÃ¶dinger equation for multiple geometries via VMC. This enables us to model continuous subsets of the potential energy surface with a single training pass. Compared to existing state-of-the-art networks, our Potential Energy Surface Network (PESNet) speeds up training for multiple geometries by up to 40 times while matching or surpassing their accuracy. This may open the path to accurate and orders of magnitude cheaper quantum mechanical calculations.

Near-Optimal Reward-Free Exploration for Linear Mixture MDPs with Plug-in Solver

- Xiaoyu Chen, Jiachen Hu, Lin Yang, Liwei Wang
- abstract@open-review(Spotlight): Although model-based reinforcement learning (RL) approaches are considered more sample efficient, existing algorithms are usually relying on sophisticated planning algorithm to couple tightly with the model-learning procedure. Hence the learned models may lack the ability of being re-used with more specialized planners. In this paper we address this issue and provide approaches to learn an RL model efficiently without the guidance of a reward signal. In particular, we take a plug-in solver approach, where we focus on learning a model in the exploration phase and demand that \emph{any planning algorithm} on the learned model can give a near-optimal policy. Specicially, we focus on the linear mixture MDP setting, where the probability transition matrix is a (unknown) convex combination of a set of existing models. We show that, by establishing a novel exploration algorithm, the plug-in approach learns a model by taking $\tilde{O}(d^2H^3/\epsilon^2)$ interactions with the environment and \emph{any} ϵ-optimal planner on the model gives an $O(\epsilon)$-optimal policy on the original model. This sample complexity matches lower bounds for non-plug-in approaches and is \emph{statistically optimal}. We achieve this result by leveraging a careful maximum total-variance bound using Bernstein inequality and properties specified to linear mixture MDP.

Meta Discovery: Learning to Discover Novel Classes given Very Limited Data

- Haoang Chi, Feng Liu, Wenjing Yang, Long Lan, Tongliang Liu, Bo Han, Gang Niu, Mingyuan Zhou, Masashi Sugiyama
- abstract@open-review(Spotlight): In novel class discovery (NCD), we are given labeled data from seen classes and unlabeled data from unseen classes, and we train clustering models for the unseen classes. However, the implicit assumptions behind NCD are still unclear. In this paper, we demystify assumptions behind NCD and find that high-level semantic features should be shared among the seen and unseen classes. Based on this finding, NCD is theoretically solvable under certain assumptions and can be naturally linked to meta-learning that has exactly the same assumption as NCD. Thus, we can empirically solve the NCD problem by meta-learning algorithms after slight modifications. This meta-learning-based methodology significantly reduces the amount of unlabeled data needed for training and makes it more practical, as demonstrated in experiments. The use of very limited data is also justified by the application scenario of NCD: since it is unnatural to label only seen-class data, NCD is sampling instead of labeling in causality. Therefore, unseen-class data should be collected on the way of collecting seen-class data, which is why they are novel and first need to be clustered.

Constrained Policy Optimization via Bayesian World Models

- Yarden As, Ilnura Usmanova, Sebastian Curi, Andreas Krause
- abstract@open-review(Spotlight): Improving sample-efficiency and safety are crucial challenges when deploying reinforcement learning in high-stakes real world applications. We propose LAMBDA, a novel model-based approach for policy optimization in safety critical tasks modeled via constrained Markov decision processes. Our approach utilizes Bayesian world models, and harnesses the resulting uncertainty to maximize optimistic upper bounds on

the task objective, as well as pessimistic upper bounds on the safety constraints. We demonstrate LAMBDA's state of the art performance on the Safety-Gym benchmark suite in terms of sample efficiency and constraint violation.

VAE Approximation Error: ELBO and Exponential Families

- Alexander Shekhovtsov, Dmitrij Schlesinger, Boris Flach
- abstract@open-review(Spotlight): The importance of Variational Autoencoders reaches far beyond standalone generative models -- the approach is also used for learning latent representations and can be generalized to semi-supervised learning. This requires a thorough analysis of their commonly known shortcomings: posterior collapse and approximation errors. This paper analyzes VAE approximation errors caused by the combination of the ELBO objective and encoder models from conditional exponential families, including, but not limited to, commonly used conditionally independent discrete and continuous models. We characterize subclasses of generative models consistent with these encoder families. We show that the ELBO optimizer is pulled away from the likelihood optimizer towards the consistent subset and study this effect experimentally. Importantly, this subset can not be enlarged, and the respective error cannot be decreased, by considering deeper encoder/decoder networks.

Generalized Decision Transformer for Offline Hindsight Information Matching

- Hiroki Furuta, Yutaka Matsuo, Shixiang Shane Gu
- abstract@open-review(Spotlight): How to extract as much learning signal from each trajectory data has been a key problem in reinforcement learning (RL), where sample inefficiency has posed serious challenges for practical applications. Recent works have shown that using expressive policy function approximators and conditioning on future trajectory information -- such as future states in hindsight experience replay (HER) or returns-to-go in Decision Transformer (DT) -- enables efficient learning of multi-task policies, where at times online RL is fully replaced by offline behavioral cloning (BC), e.g. sequence modeling. We demonstrate that all these approaches are doing hindsight information matching (HIM) -- training policies that can output the rest of trajectory that matches some statistics of future state information. We present Generalized Decision Transformer (GDT) for solving any HIM problem, and show how different choices for the feature function and the anti-causal aggregator not only recover DT as a special case, but also lead to novel Categorical DT (CDT) and Bi-directional DT (BDT) for matching different statistics of the future. For evaluating CDT and BDT, we define offline multi-task state-marginal matching (SMM) and imitation learning (IL) as two generic HIM problems, propose a Wasserstein distance loss as a metric for both, and empirically study them on MuJoCo continuous control benchmarks. Categorical DT, which simply replaces anti-causal summation with anti-causal binning in DT, enables arguably the first effective offline multi-task SMM algorithm that generalizes well to unseen (and even synthetic) multi-modal reward or state-feature distributions. Bi-directional DT, which uses an anti-causal second transformer as the aggregator, can learn to model any statistics of the future and outperforms DT variants in offline multi-task IL, i.e. one-shot IL. Our generalized formulations from HIM and GDT greatly expand the role of powerful sequence modeling architectures in modern RL.

Unifying Likelihood-free Inference with Black-box Optimization and Beyond

- Dinghuai Zhang, Jie Fu, Yoshua Bengio, Aaron Courville
- abstract@open-review(Spotlight): Black-box optimization formulations for biological sequence design have drawn recent attention due to their promising potential impact on the pharmaceutical industry. In this work, we propose to unify two seemingly distinct worlds: likelihood-free inference and black-box optimization, under one probabilistic framework. In tandem, we provide a recipe for constructing various sequence design methods based on this framework. We show how previous optimization approaches can be "reinvented" in our framework, and further propose new probabilistic black-box optimization algorithms. Extensive experiments on sequence design application illustrate the benefits of the proposed methodology.

DEPTS: Deep Expansion Learning for Periodic Time Series Forecasting

- Wei Fan, Shun Zheng, Xiaohan Yi, Wei Cao, Yanjie Fu, Jiang Bian, Tie-Yan Liu
- abstract@open-review(Spotlight): Periodic time series (PTS) forecasting plays a crucial role in a variety of industries to foster critical tasks, such as early warning, pre-planning, resource scheduling, etc. However, the complicated dependencies of the PTS signal on its inherent periodicity as well as the sophisticated composition of various periods hinder the performance of PTS forecasting. In this paper, we introduce a deep expansion learning framework, DEPTS, for PTS forecasting. DEPTS starts with a decoupled formulation by introducing the periodic state as a hidden variable, which stimulates us to make two dedicated modules to tackle the aforementioned two challenges. First, we develop an expansion module on top of residual learning to perform a layer-by-layer expansion of those complicated dependencies. Second, we introduce a periodicity module with a parameterized periodic function that holds sufficient capacity to capture diversified periods. Moreover, our two customized modules also have certain interpretable capabilities, such as attributing the forecasts to either local momenta or global periodicity and characterizing certain core periodic properties, e.g., amplitudes and frequencies. Extensive experiments on both synthetic data and real-world data demonstrate the effectiveness of DEPTS on handling PTS. In most cases, DEPTS achieves significant improvements over the best baseline. Specifically, the error reduction can even reach up to 20% for a few cases. All codes for this paper are publicly available.

Evaluation Metrics for Graph Generative Models: Problems, Pitfalls, and Practical Solutions

- Leslie O'Bray, Max Horn, Bastian Rieck, Karsten Borgwardt
- abstract@open-review(Spotlight): Graph generative models are a highly active branch of machine learning. Given the steady development of new models of ever-increasing complexity, it is necessary to provide a principled way to evaluate and compare them. In this paper, we enumerate the desirable criteria for such a comparison metric and provide an overview of the status quo of graph generative model comparison in use today, which predominantly relies on the maximum mean discrepancy (MMD). We perform a systematic evaluation of MMD in the context of graph generative model comparison, highlighting some of the challenges and pitfalls researchers inadvertently may encounter. After conducting a thorough analysis of the behaviour of MMD on synthetically-generated perturbed graphs as well as on recently-proposed graph generative models, we are able to provide a suitable procedure to mitigate these challenges and pitfalls. We aggregate our findings into a list of practical recommendations for researchers to use when evaluating graph generative models.

Natural Posterior Network: Deep Bayesian Predictive Uncertainty for Exponential Family Distributions

- Bertrand Charpentier, Oliver Borchert, Daniel ZÃ¼gner, Simon Geisler, Stephan GÃ¼nnemann
- abstract@open-review(Spotlight): Uncertainty awareness is crucial to develop reliable machine learning models. In this work, we propose the Natural Posterior Network (NatPN) for fast and high-quality uncertainty estimation for any task where the target distribution belongs to the exponential family. Thus, NatPN finds application for both classification and general regression settings. Unlike many previous approaches, NatPN does not require out-of-distribution (OOD) data at training time. Instead, it leverages Normalizing Flows to fit a single density on a learned low-dimensional and task-dependent latent space. For any input sample, NatPN uses the predicted likelihood to perform a Bayesian update over the target distribution. Theoretically, NatPN assigns high uncertainty far away from training data. Empirically, our extensive experiments on calibration and OOD detection show that NatPN delivers highly competitive performance for classification, regression and count prediction tasks.

Learning Altruistic Behaviours in Reinforcement Learning without External Rewards

- Tim Franzmeyer, Mateusz Malinowski, Joao F. Henriques
- abstract@open-review(Spotlight): Can artificial agents learn to assist others in achieving their goals without knowing what those goals are? Generic reinforcement learning agents could be trained to behave altruistically towards others by rewarding them for altruistic behaviour, i.e., rewarding them for benefiting other agents in a given situation. Such an approach assumes that other agents' goals are known so that the altruistic agent can cooperate in achieving those goals. However, explicit knowledge of other agents' goals is often difficult to acquire. In the case of human agents, their goals and preferences may be difficult to express fully; they might be ambiguous or even contradictory. Thus, it is beneficial to develop agents that do not depend on external supervision and learn altruistic behaviour in a task-agnostic manner. We propose to act altruistically towards other agents by giving them more choice and allowing them to achieve their goals better. Some concrete examples include opening a door for others or safeguarding them to pursue their objectives without interference. We formalize this concept and propose an altruistic agent that learns to increase the choices another agent has by preferring to maximize the number of states that the other agent can reach in its future. We evaluate our approach in three different multi-agent environments where another agent's success depends on altruistic behaviour. Finally, we show that our unsupervised agents can perform comparably to agents explicitly trained to work cooperatively, in some cases even outperforming them.

Context-Aware Sparse Deep Coordination Graphs

- Tonghan Wang, Liang Zeng, Weijun Dong, Qianlan Yang, Yang Yu, Chongjie Zhang
- abstract@open-review(Spotlight): Learning sparse coordination graphs adaptive to the coordination dynamics among agents is a long-standing problem in cooperative multi-agent learning. This paper studies this problem and proposes a novel method using the variance of payoff functions to construct context-aware sparse coordination topologies. We theoretically consolidate our method by proving that the smaller the variance of payoff functions is, the less likely action selection will change after removing the corresponding edge. Moreover, we propose to learn action representations to effectively reduce the influence of payoff functions' estimation errors on graph construction. To empirically evaluate our method, we present the Multi-Agent COordination (MACO) benchmark by collecting classic coordination problems in the literature, increasing their difficulty, and classifying them into different types. We carry out a case study and experiments on the MACO and StarCraft II micromanagement benchmark to demonstrate the dynamics of sparse graph learning, the influence of graph sparseness, and the learning performance of our method.

On the approximation properties of recurrent encoder-decoder architectures

- Zhong Li, Haotian Jiang, Qianxiao Li
- abstract@open-review(Spotlight): Encoder-decoder architectures have recently gained popularity in sequence to sequence modelling, featuring in state-of-the-art models such as transformers. However, a mathematical understanding of their working principles still remains limited. In this paper, we study the approximation properties of recurrent encoder-decoder architectures. Prior work established theoretical results for RNNs in the linear setting, where approximation capabilities can be related to smoothness and memory of target temporal relationships. Here, we uncover that the encoder and decoder together form a particular â€œtemporal product structureâ€ which determines the approximation efficiency. Moreover, the encoder-decoder architecture generalises RNNs with the capability to learn time-inhomogeneous relationships. Our results provide the theoretical understanding of approximation properties of the recurrent encoder-decoder architecture, which precisely characterises, in the considered setting, the types of temporal relationships that can be efficiently learned.

Pixelated Butterfly: Simple and Efficient Sparse training for Neural Network Models

- Beidi Chen, Tri Dao, Kaizhao Liang, Jiaming Yang, Zhao Song, Atri Rudra, Christopher Re
- abstract@open-review(Spotlight): Overparameterized neural networks generalize well but are expensive to train. Ideally one would like to reduce their computational cost while retaining their generalization benefits. Sparse model training is a simple and promising approach to achieve this, but there remain challenges as existing methods struggle with accuracy loss, slow training runtime, or difficulty in sparsifying all model components. The core problem is that searching for a sparsity mask over a discrete set of sparse matrices is difficult and expensive. To address this, our main insight is to optimize over a continuous superset of sparse matrices with a fixed structure known as products of butterfly matrices. As butterfly matrices are not hardware efficient, we propose simple variants of butterfly (block and flat) to take advantage of modern hardware. Our method (Pixelated Butterfly) uses a simple fixed sparsity pattern based on flat block butterfly and low-rank matrices to sparsify most network layers (e.g., attention, MLP). We empirically validate that Pixelated Butterfly is $3\times$ faster than Butterfly and speeds up training to achieve favorable accuracy--efficiency tradeoffs. On the ImageNet classification and WikiText-103 language modeling tasks, our sparse models train up to 2.3$\times$ faster than the dense MLP-Mixer, Vision Transformer, and GPT-2 small with no drop in accuracy.

8-bit Optimizers via Block-wise Quantization

- Tim Dettmers, Mike Lewis, Sam Shleifer, Luke Zettlemoyer
- abstract@open-review(Spotlight): Stateful optimizers maintain gradient statistics over time, e.g., the exponentially smoothed sum (SGD with momentum) or squared sum (Adam) of past gradient values. This state can be used to accelerate optimization significantly, compared to plain stochastic gradient descent, but uses memory that might otherwise be allocated to model parameters, thereby limiting the maximum size of models trained in practice. In this paper, we develop the first optimizers that use 8-bit statistics while maintaining the performance levels of using 32-bit optimizer states. To overcome the resulting computational, quantization, and stability challenges, we develop block-wise dynamic quantization. Block-wise quantization divides input tensors into smaller blocks that are independently quantized. Each block is processed in parallel across cores, yielding faster optimization and high precision quantization. To maintain stability and performance, we combine block-wise quantization with two additional changes: (1) dynamic quantization, a form of non-linear optimization that is precise for both large and small magnitude values, and (2) a stable embedding layer to reduce gradient variance that comes from the highly non-uniform distribution of input tokens in language models. As a result, our 8-bit optimizers maintain 32-bit performance with a small fraction of the memory footprint on a range of tasks, including 1.5B parameter language modeling, GLUE finetuning, ImageNet classification, WMT'14 machine translation, MoCo v2 contrastive ImageNet pretraining+finetuning, and RoBERTa pretraining, without changes to the original optimizer hyperparameters. We open-source our 8-bit optimizers as a drop-in replacement that only requires a two-line code change.

Finding Biological Plausibility for Adversarially Robust Features via Metameric Tasks

- Anne Harrington, Arturo Deza
- abstract@open-review(Spotlight): Recent work suggests that feature constraints in the training datasets of deep neural networks (DNNs) drive robustness to adversarial noise (Ilyas et al., 2019). The representations learned by such adversarially robust networks have also been shown to be more human perceptually-aligned than non-robust networks via image manipulations (Santurkar et al., 2019, Engstrom et al., 2019). Despite appearing closer to human visual perception, it is unclear if the constraints in robust DNN representations match biological constraints found in human vision. Human vision seems to rely on texture-based/summary statistic representations in the periphery, which have been shown to explain phenomena such as crowding (Balas et al., 2009) and performance on visual search tasks (Rosenholtz et al., 2012). To understand how adversarially robust optimizations/representations compare to human vision, we performed a psychophysics experiment using a metamer task similar to Freeman \& Simoncelli, 2011, Wallis et al., 2016 and Deza et al., 2019 where we evaluated how well human observers could distinguish between images synthesized to match adversarially robust representations compared to non-robust representations and a texture synthesis model of peripheral vision (Texforms a la Long et al., 2018). We found that the discriminability of robust representation and texture model images decreased to near chance performance as stimuli were presented farther in the periphery. Moreover, performance on robust and texture-model images showed similar trends within participants, while performance on non-robust representations changed minimally across the visual field. These results together suggest that (1) adversarially robust representations capture peripheral

computation better than non-robust representations and (2) robust representations capture peripheral computation similar to current state-of-the-art texture peripheral vision models. More broadly, our findings support the idea that localized texture summary statistic representations may drive human invariance to adversarial perturbations and that the incorporation of such representations in DNNs could give rise to useful properties like adversarial robustness.

Omni-Dimensional Dynamic Convolution

- Chao Li, Aojun Zhou, Anbang Yao
- abstract@open-review(Spotlight): Learning a single static convolutional kernel in each convolutional layer is the common training paradigm of modern Convolutional Neural Networks (CNNs). Instead, recent research in dynamic convolution shows that learning a linear combination of n convolutional kernels weighted with their input-dependent attentions can significantly improve the accuracy of light-weight CNNs, while maintaining efficient inference. However, we observe that existing works endow convolutional kernels with the dynamic property through one dimension (regarding the convolutional kernel number) of the kernel space, but the other three dimensions (regarding the spatial size, the input channel number and the output channel number for each convolutional kernel) are overlooked. Inspired by this, we present Omni-dimensional Dynamic Convolution (ODConv), a more generalized yet elegant dynamic convolution design, to advance this line of research. ODConv leverages a novel multi-dimensional attention mechanism with a parallel strategy to learn complementary attentions for convolutional kernels along all four dimensions of the kernel space at any convolutional layer. As a drop-in replacement of regular convolutions, ODConv can be plugged into many CNN architectures. Extensive experiments on the ImageNet and MS-COCO datasets show that ODConv brings solid accuracy boosts for various prevailing CNN backbones including both light-weight and large ones, e.g., 3.77%~5.71%|1.86%~3.72% absolute top-1 improvements to MobivleNetV2|ResNet family on the ImageNet dataset. Intriguingly, thanks to its improved feature learning ability, ODConv with even one single kernel can compete with or outperform existing dynamic convolution counterparts with multiple kernels, substantially reducing extra parameters. Furthermore, ODConv is also superior to other attention modules for modulating the output features or the convolutional weights. Code and models will be available at https://github.com/OSVAI/ODConv.

EViT: Expediting Vision Transformers via Token Reorganizations

- Youwei Liang, Chongjian GE, Zhan Tong, Yibing Song, Jue Wang, Pengtao Xie
- abstract@open-review(Spotlight): Vision Transformers (ViTs) take all the image patches as tokens and construct multi-head self-attention (MHSA) among them. Complete leverage of these image tokens brings redundant computations since not all the tokens are attentive in MHSA. Examples include that tokens containing semantically meaningless or distractive image backgrounds do not positively contribute to the ViT predictions. In this work, we propose to reorganize image tokens during the feed-forward process of ViT models, which is integrated into ViT during training. For each forward inference, we identify the attentive image tokens between MHSA and FFN (i.e., feed-forward network) modules, which is guided by the corresponding class token attention. Then, we reorganize image tokens by preserving attentive image tokens and fusing inattentive ones to expedite subsequent MHSA and FFN computations. To this end, our method EViT improves ViTs from two perspectives. First, under the same amount of input image tokens, our method reduces MHSA and FFN computation for efficient inference. For instance, the inference speed of DeiT-S is increased by 50% while its recognition accuracy is decreased by only 0.3% for ImageNet classification. Second, by maintaining the same computational cost, our method empowers ViTs to take more image tokens as input for recognition accuracy improvement, where the image tokens are from higher resolution images. An example is that we improve the recognition accuracy of DeiT-S by 1% for ImageNet classification at the same computational cost of a vanilla DeiT-S. Meanwhile, our method does not introduce more parameters to ViTs. Experiments on the standard benchmarks show the effectiveness of our method. The code is available at https://github.com/youweiliang/evit

D-CODE: Discovering Closed-form ODEs from Observed Trajectories

- Zhaozhi Qian, Krzysztof Kacprzyk, Mihaela van der Schaar
- abstract@open-review(Spotlight): For centuries, scientists have manually designed closed-form ordinary differential equations (ODEs) to model dynamical systems. An automated tool to distill closed-form ODEs from observed trajectories would accelerate the modeling process. Traditionally, symbolic regression is used to uncover a closed-form prediction function $a=f(b)$ with label-feature pairs (a_i, b_i) as training examples. However, an ODE models the time derivative $\dot{x}(t)$ of a dynamical system, e.g. $\dot{x}(t) = f(x(t),t)$, and the "label" $\dot{x}(t)$ is usually *not* observed. The existing ways to bridge this gap only perform well for a narrow range of settings with low measurement noise, frequent sampling, and non-chaotic dynamics. In this work, we propose the Discovery of Closed-form ODE framework (D-CODE), which advances symbolic regression beyond the paradigm of supervised learning. D-CODE leverages a novel objective function based on the variational formulation of ODEs to bypass the unobserved time derivative. For formal justification, we prove that this objective is a valid proxy for the estimation error of the true (but unknown) ODE. In the experiments, D-CODE successfully discovered the governing equations of a diverse range of dynamical systems under challenging measurement settings with high noise and infrequent sampling.

Spanning Tree-based Graph Generation for Molecules

- Sungsoo Ahn, Binghong Chen, Tianzhe Wang, Le Song
- abstract@open-review(Spotlight): In this paper, we explore the problem of generating molecules using deep neural networks, which has recently gained much interest in chemistry. To this end, we propose a spanning tree-based graph generation (STGG) framework based on formulating molecular graph generation as a construction of a spanning tree and the residual edges. Such a formulation exploits the sparsity of molecular graphs and allows using compact tree-constructive operations to define the molecular graph connectivity. Based on the intermediate graph structure of the construction process, our framework can constrain its generation to molecular graphs that satisfy the chemical valence rules. We also newly design a Transformer architecture with tree-based relative positional encodings for realizing the tree construction procedure. Experiments on QM9, ZINC250k, and MOSES benchmarks verify the effectiveness of the proposed framework in metrics such as validity, Frechet ChemNet distance, and fragment similarity. We also demonstrate the usefulness of STGG in maximizing penalized LogP value of molecules.

Policy improvement by planning with Gumbel

- Ivo Danihelka, Arthur Guez, Julian Schrittwieser, David Silver
- abstract@open-review(Spotlight): AlphaZero is a powerful reinforcement learning algorithm based on approximate policy iteration and tree search. However, AlphaZero can fail to improve its policy network, if not visiting all actions at the root of a search tree. To address this issue, we propose a policy improvement algorithm based on sampling actions without replacement. Furthermore, we use the idea of policy improvement to replace the more heuristic mechanisms by which AlphaZero selects and uses actions, both at root nodes and at non-root nodes. Our new algorithms, Gumbel AlphaZero and Gumbel MuZero, respectively without and with model-learning, match the state of the art on Go, chess, and Atari, and significantly improve prior performance when planning with few simulations.

Learning Optimal Conformal Classifiers

- David Stutz, Krishnamurthy Dj Dvijotham, Ali Taylan Cemgil, Arnaud Doucet
- abstract@open-review(Spotlight): Modern deep learning based classifiers show very high accuracy on test data but this does not provide sufficient guarantees for safe deployment, especially in high-stake AI applications such as medical diagnosis. Usually, predictions are obtained without a reliable uncertainty estimate or a formal guarantee. Conformal prediction (CP) addresses these issues by using the classifier's predictions, e.g., its probability estimates, to predict confidence sets containing the true class with a user-specified probability. However, using CP as a separate processing step after

training prevents the underlying model from adapting to the prediction of confidence sets. Thus, this paper explores strategies to differentiate through CP during training with the goal of training model with the conformal wrapper end-to-end. In our approach, conformal training (ConfTr), we specifically "simulate" conformalization on mini-batches during training. Compared to standard training, ConfTr reduces the average confidence set size (inefficiency) of state-of-the-art CP methods applied after training. Moreover, it allows to "shape" the confidence sets predicted at test time, which is difficult for standard CP. On experiments with several datasets, we show ConfTr can influence how inefficiency is distributed across classes, or guide the composition of confidence sets in terms of the included classes, while retaining the guarantees offered by CP.

Multitask Prompted Training Enables Zero-Shot Task Generalization

- Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, Alexander M Rush
- abstract@open-review(Spotlight): Large language models have recently been shown to attain reasonable zero-shot generalization on a diverse set of tasks (Brown et al., 2020). It has been hypothesized that this is a consequence of implicit multitask learning in language modelsâ€™ pretraining (Radford et al., 2019). Can zero-shot generalization instead be directly induced by explicit multitask learning? To test this question at scale, we develop a system for easily mapping any natural language tasks into a human-readable prompted form. We convert a large set of supervised datasets, each with multiple prompts with diverse wording. These prompted datasets allow for benchmarking the ability of a model to perform completely unseen tasks specified in natural language. We fine-tune a pretrained encoder-decoder model (Raffel et al., 2020; Lester et al., 2021) on this multitask mixture covering a wide variety of tasks. The model attains strong zero-shot performance on several datasets, often outperforming models 16Ã— its size. Further, our model attains strong performance on a subset of tasks from the BIG-Bench benchmark, outperforming models 6Ã— its size. All trained models are available at https://github.com/bigscience-workshop/t-zero, and all prompts are available at https://github.com/bigscience-workshop/promptsource.

Continuous-Time Meta-Learning with Forward Mode Differentiation

- Tristan Deleu, David Kanaa, Leo Feng, Giancarlo Kerg, Yoshua Bengio, Guillaume Lajoie, Pierre-Luc Bacon
- abstract@open-review(Spotlight): Drawing inspiration from gradient-based meta-learning methods with infinitely small gradient steps, we introduce Continuous-Time Meta-Learning (COMLN), a meta-learning algorithm where adaptation follows the dynamics of a gradient vector field. Specifically, representations of the inputs are meta-learned such that a task-specific linear classifier is obtained as a solution of an ordinary differential equation (ODE). Treating the learning process as an ODE offers the notable advantage that the length of the trajectory is now continuous, as opposed to a fixed and discrete number of gradient steps. As a consequence, we can optimize the amount of adaptation necessary to solve a new task using stochastic gradient descent, in addition to learning the initial conditions as is standard practice in gradient-based meta-learning. Importantly, in order to compute the exact meta-gradients required for the outer-loop updates, we devise an efficient algorithm based on forward mode differentiation, whose memory requirements do not scale with the length of the learning trajectory, thus allowing longer adaptation in constant memory. We provide analytical guarantees for the stability of COMLN, we show empirically its efficiency in terms of runtime and memory usage, and we illustrate its effectiveness on a range of few-shot image classification problems.

On the relation between statistical learning and perceptual distances

- Alexander Hepburn, Valero Laparra, Raul Santos-Rodriguez, Johannes BallÃ©, Jesus Malo
- abstract@open-review(Spotlight): It has been demonstrated many times that the behavior of the human visual system is connected to the statistics of natural images. Since machine learning relies on the statistics of training data as well, the above connection has interesting implications when using perceptual distances (which mimic the behavior of the human visual system) as a loss function. In this paper, we aim to unravel the non-trivial relationships between the probability distribution of the data, perceptual distances, and unsupervised machine learning. To this end, we show that perceptual sensitivity is correlated with the probability of an image in its close neighborhood. We also explore the relation between distances induced by autoencoders and the probability distribution of the training data, as well as how these induced distances are correlated with human perception. Finally, we find perceptual distances do not always lead to noticeable gains in performance over Euclidean distance in common image processing tasks, except when data is scarce and the perceptual distance provides regularization. We propose this may be due to a double-counting effect of the image statistics, once in the perceptual distance and once in the training procedure.

Implicit Bias of Projected Subgradient Method Gives Provable Robust Recovery of Subspaces of Unknown Codimension

- Paris Giampouras, Benjamin David Haeffele, Rene Vidal
- abstract@open-review(Spotlight): Robust subspace recovery (RSR) is the problem of learning a subspace from sample data points corrupted by outliers. Dual Principal Component Pursuit (DPCP) is a robust subspace recovery method that aims to find a basis for the orthogonal complement of the subspace by minimizing the sum of the distances of the points to the subspaces subject to orthogonality constraints on the basis. Prior work has shown that DPCP can provably recover the correct subspace in the presence of outliers as long as the true dimension of the subspace is known. In this paper, we show that if the orthogonality constraints --adopted in previous DPCP formulations-- are relaxed and random initialization is used instead of spectral one, DPCP can provably recover a subspace of \emph{unknown dimension}. Specifically, we propose a very simple algorithm based on running multiple instances of a projected sub-gradient descent method (PSGM), with each problem instance seeking to find one vector in the null space of the subspace. We theoretically prove that under mild conditions this approach succeeds with high probability. In particular, we show that 1) all of the problem instances will converge to a vector in the nullspace of the subspace and 2) the ensemble of problem instance solutions will be sufficiently diverse to fully span the nullspace of the subspace thus also revealing its true unknown codimension. We provide empirical results that corroborate our theoretical results and showcase the remarkable implicit rank regularization behavior of the PSGM algorithm that allows us to perform RSR without knowing the subspace dimension

On the Optimal Memorization Power of ReLU Neural Networks

- Gal Vardi, Gilad Yehudai, Ohad Shamir
- abstract@open-review(Spotlight): We study the memorization power of feedforward ReLU neural networks. We show that such networks can memorize any N points that satisfy a mild separability assumption using $\tilde{O}\left(\sqrt{N}\right)$ parameters. Known VC-dimension upper bounds imply that memorizing N samples requires $\Omega(\sqrt{N})$ parameters, and hence our construction is optimal up to logarithmic factors. We also give a generalized construction for networks with depth bounded by $1 \leq L \leq \sqrt{N}$, for memorizing N samples using $\tilde{O}(N/L)$ parameters. This bound is also optimal up to logarithmic factors. Our construction uses weights with large bit complexity. We prove that having such a large bit complexity is both necessary and sufficient for memorization with a sub-linear number of parameters.

Programmatic Reinforcement Learning without Oracles

- Wenjie Qiu, He Zhu
- abstract@open-review(Spotlight): Deep reinforcement learning (RL) has led to encouraging successes in many challenging control tasks. However, a deep RL model lacks interpretability due to the difficulty of identifying how the model's control logic relates to its network structure. Programmatic policies structured in more interpretable representations emerge as a promising solution. Yet two shortcomings remain: First, synthesizing programmatic policies

requires optimizing over the discrete and non-differentiable search space of program architectures. Previous works are suboptimal because they only enumerate program architectures greedily guided by a pretrained RL oracle. Second, these works do not exploit compositionality, an important programming concept, to reuse and compose primitive functions to form a complex function for new tasks. Our first contribution is a programmatically interpretable RL framework that conducts program architecture search on top of a continuous relaxation of the architecture space defined by programming language grammar rules. Our algorithm allows policy architectures to be learned with policy parameters via bilevel optimization using efficient policy-gradient methods, and thus does not require a pretrained oracle. Our second contribution is improving programmatic policies to support compositionality by integrating primitive functions learned to grasp task-agnostic skills as a composite program to solve novel RL problems. Experiment results demonstrate that our algorithm excels in discovering optimal programmatic policies that are highly interpretable. The code of this work is available at https://github.com/RU-Automated-Reasoning-Group/pi-PRL.

When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations

- Xiangning Chen, Cho-Jui Hsieh, Boqing Gong
- abstract@open-review(Spotlight): Vision Transformers (ViTs) and MLPs signal further efforts on replacing hand-wired features or inductive biases with general-purpose neural architectures. Existing works empower the models by massive data, such as large-scale pre-training and/or repeated strong data augmentations, and still report optimization-related problems (e.g., sensitivity to initialization and learning rates). Hence, this paper investigates ViTs and MLP-Mixers from the lens of loss geometry, intending to improve the models' data efficiency at training and generalization at inference. Visualization and Hessian reveal extremely sharp local minima of converged models. By promoting smoothness with a recently proposed sharpness-aware optimizer, we substantially improve the accuracy and robustness of ViTs and MLP-Mixers on various tasks spanning supervised, adversarial, contrastive, and transfer learning (e.g., +5.3\% and +11.0\% top-1 accuracy on ImageNet for ViT-B/16 and Mixer-B/16, respectively, with the simple Inception-style preprocessing). We show that the improved smoothness attributes to sparser active neurons in the first few layers. The resultant ViTs outperform ResNets of similar size and throughput when trained from scratch on ImageNet without large-scale pre-training or strong data augmentations. Model checkpoints are available at \url{https://github.com/google-research/vision_transformer}.

Online Hyperparameter Meta-Learning with Hypergradient Distillation

- Hae Beom Lee, Hayeon Lee, JaeWoong Shin, Eunho Yang, Timothy Hospedales, Sung Ju Hwang
- abstract@open-review(Spotlight): Many gradient-based meta-learning methods assume a set of parameters that do not participate in inner-optimization, which can be considered as hyperparameters. Although such hyperparameters can be optimized using the existing gradient-based hyperparameter optimization (HO) methods, they suffer from the following issues. Unrolled differentiation methods do not scale well to high-dimensional hyperparameters or horizon length, Implicit Function Theorem (IFT) based methods are restrictive for online optimization, and short horizon approximations suffer from short horizon bias. In this work, we propose a novel HO method that can overcome these limitations, by approximating the second-order term with knowledge distillation. Specifically, we parameterize a single Jacobian-vector product (JVP) for each HO step and minimize the distance from the true second-order term. Our method allows online optimization and also is scalable to the hyperparameter dimension and the horizon length. We demonstrate the effectiveness of our method on three different meta-learning methods and two benchmark datasets.

Tighter Sparse Approximation Bounds for ReLU Neural Networks

- Carles Domingo-Enrich, Youssef Mroueh
- abstract@open-review(Spotlight): A well-known line of work (Barron, 1993; Breiman, 1993; Klusowski & Barron, 2018) provides bounds on the width n of a ReLU two-layer neural network needed to approximate a function f over the ball $\mathcal{B}*R(\mathbb{R}^d)$ up to error ϵ, when the Fourier based quantity $C_f = \int{\mathbb{R}^d} \|\xi\|^2 |\hat{f}(\xi)| \ d\xi$ is finite. More recently Ongie et al. (2019) used the Radon transform as a tool for analysis of infinite-width ReLU two-layer networks. In particular, they introduce the concept of Radon-based $\mathcal{R}$-norms and show that a function defined on $\mathbb{R}^d$ can be represented as an infinite-width two-layer neural network if and only if its $\mathcal{R}$-norm is finite. In this work, we extend the framework of Ongie et al. (2019) and define similar Radon-based semi-norms ($\mathcal{R}, \mathcal{U}$-norms) such that a function admits an infinite-width neural network representation on a bounded open set $\mathcal{U} \subseteq \mathbb{R}^d$ when its $\mathcal{R}, \mathcal{U}$-norm is finite. Building on this, we derive sparse (finite-width) neural network approximation bounds that refine those of Breiman (1993); Klusowski & Barron (2018). Finally, we show that infinite-width neural network representations on bounded open sets are not unique and study their structure, providing a functional view of mode connectivity.

Long Expressive Memory for Sequence Modeling

- T. Konstantin Rusch, Siddhartha Mishra, N. Benjamin Erichson, Michael W. Mahoney
- abstract@open-review(Spotlight): We propose a novel method called Long Expressive Memory (LEM) for learning long-term sequential dependencies. LEM is gradient-based, it can efficiently process sequential tasks with very long-term dependencies, and it is sufficiently expressive to be able to learn complicated input-output maps. To derive LEM, we consider a system of multiscale ordinary differential equations, as well as a suitable time-discretization of this system. For LEM, we derive rigorous bounds to show the mitigation of the exploding and vanishing gradients problem, a well-known challenge for gradient-based recurrent sequential learning methods. We also prove that LEM can approximate a large class of dynamical systems to high accuracy. Our empirical results, ranging from image and time-series classification through dynamical systems prediction to speech recognition and language modeling, demonstrate that LEM outperforms state-of-the-art recurrent neural networks, gated recurrent units, and long short-term memory models.

Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy

- Jiehui Xu, Haixu Wu, Jianmin Wang, Mingsheng Long
- abstract@open-review(Spotlight): Unsupervised detection of anomaly points in time series is a challenging problem, which requires the model to derive a distinguishable criterion. Previous methods tackle the problem mainly through learning pointwise representation or pairwise association, however, neither is sufficient to reason about the intricate dynamics. Recently, Transformers have shown great power in unified modeling of pointwise representation and pairwise association, and we find that the self-attention weight distribution of each time point can embody rich association with the whole series. Our key observation is that due to the rarity of anomalies, it is extremely difficult to build nontrivial associations from abnormal points to the whole series, thereby, the anomalies' associations shall mainly concentrate on their adjacent time points. This adjacent-concentration bias implies an association-based criterion inherently distinguishable between normal and abnormal points, which we highlight through the Association Discrepancy. Technically, we propose the Anomaly Transformer with a new Anomaly-Attention mechanism to compute the association discrepancy. A minimax strategy is devised to amplify the normal-abnormal distinguishability of the association discrepancy. The Anomaly Transformer achieves state-of-the-art results on six unsupervised time series anomaly detection benchmarks of three applications: service monitoring, space & earth exploration, and water treatment.

Progressive Distillation for Fast Sampling of Diffusion Models

- Tim Salimans, Jonathan Ho
- abstract@open-review(Spotlight): Diffusion models have recently shown great promise for generative modeling, outperforming GANs on perceptual quality and autoregressive models at density estimation. A remaining downside is their slow sampling time: generating high quality samples takes many hundreds or thousands of model evaluations. Here we make two contributions to help eliminate this downside: First, we present new parameterizations of

diffusion models that provide increased stability when using few sampling steps, compared to models in the literature. Second, we present a method to distill a trained deterministic diffusion sampler, using many steps, into a new diffusion model that takes half as many sampling steps. We then keep progressively applying this distillation procedure to our model, halving the number of required sampling steps each time. On standard image generation benchmarks like CIFAR-10, ImageNet, and LSUN, we start out with (near) state-of-the-art samplers taking 1024 or 8192 steps, and are able to distill down to models taking as little as 4 steps without losing much perceptual quality; achieving, for example, a FID of 3.0 on CIFAR-10 in 4 steps. Finally, we show that the full progressive distillation procedure does not take more time than it takes to train the original model, thus representing an efficient solution for generative modeling using diffusion at both train and test time.

A General Analysis of Example-Selection for Stochastic Gradient Descent

- Yucheng Lu, Si Yi Meng, Christopher De Sa
- abstract@open-review(Spotlight): Training example order in SGD has long been known to affect convergence rate. Recent results show that accelerated rates are possible in a variety of cases for permutation-based sample orders, in which each example from the training set is used once before any example is reused. In this paper, we develop a broad condition on the sequence of examples used by SGD that is sufficient to prove tight convergence rates in both strongly convex and non-convex settings. We show that our approach suffices to recover, and in some cases improve upon, previous state-of-the-art analyses for four known example-selection schemes: (1) shuffle once, (2) random reshuffling, (3) random reshuffling with data echoing, and (4) Markov Chain Gradient Descent. Motivated by our theory, we propose two new example-selection approaches. First, using quasi-Monte-Carlo methods, we achieve unprecedented accelerated convergence rates for learning with data augmentation. Second, we greedily choose a fixed scan-order to minimize the metric used in our condition and show that we can obtain more accurate solutions from the same number of epochs of SGD. We conclude by empirically demonstrating the utility of our approach for both convex linear-model and deep learning tasks. Our code is available at: https://github.com/EugeneLYC/qmc-ordering.

Assessing Generalization of SGD via Disagreement

- Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, J Zico Kolter
- abstract@open-review(Spotlight): We empirically show that the test error of deep networks can be estimated by training the same architecture on the same training set but with two different runs of Stochastic Gradient Descent (SGD), and then measuring the disagreement rate between the two networks on unlabeled test data. This builds on -- and is a stronger version of -- the observation in Nakkiran&Bansal 20, which requires the runs to be on separate training sets. We further theoretically show that this peculiar phenomenon arises from the well-calibrated nature of ensembles of SGD-trained models. This finding not only provides a simple empirical measure to directly predict the test error using unlabeled test data, but also establishes a new conceptual connection between generalization and calibration.

Generative Planning for Temporally Coordinated Exploration in Reinforcement Learning

- Haichao Zhang, Wei Xu, Haonan Yu
- abstract@open-review(Spotlight): Standard model-free reinforcement learning algorithms optimize a policy that generates the action to be taken in the current time step in order to maximize expected future return. While flexible, it faces difficulties arising from the inefficient exploration due to its single step nature. In this work, we present Generative Planning method (GPM), which can generate actions not only for the current step, but also for a number of future steps (thus termed as generative planning). This brings several benefits to GPM. Firstly, since GPM is trained by maximizing value, the plans generated from it can be regarded as intentional action sequences for reaching high value regions. GPM can therefore leverage its generated multi-step plans for temporally coordinated exploration towards high value regions, which is potentially more effective than a sequence of actions generated by perturbing each action at single step level, whose consistent movement decays exponentially with the number of exploration steps. Secondly, starting from a crude initial plan generator, GPM can refine it to be adaptive to the task, which, in return, benefits future explorations. This is potentially more effective than commonly used action-repeat strategy, which is non-adaptive in its form of plans. Additionally, since the multi-step plan can be interpreted as the intent of the agent from now to a span of time period into the future, it offers a more informative and intuitive signal for interpretation. Experiments are conducted on several benchmark environments and the results demonstrated its effectiveness compared with several baseline methods.

Pessimistic Bootstrapping for Uncertainty-Driven Offline Reinforcement Learning

- Chenjia Bai, Lingxiao Wang, Zhuoran Yang, Zhi-Hong Deng, Animesh Garg, Peng Liu, Zhaoran Wang
- abstract@open-review(Spotlight): Offline Reinforcement Learning (RL) aims to learn policies from previously collected datasets without exploring the environment. Directly applying off-policy algorithms to offline RL usually fails due to the extrapolation error caused by the out-of-distribution (OOD) actions. Previous methods tackle such problem by penalizing the Q-values of OOD actions or constraining the trained policy to be close to the behavior policy. Nevertheless, such methods typically prevent the generalization of value functions beyond the offline data and also lack precise characterization of OOD data. In this paper, we propose Pessimistic Bootstrapping for offline RL (PBRL), a purely uncertainty-driven offline algorithm without explicit policy constraints. Specifically, PBRL conducts uncertainty quantification via the disagreement of bootstrapped Q-functions, and performs pessimistic updates by penalizing the value function based on the estimated uncertainty. To tackle the extrapolating error, we further propose a novel OOD sampling method. We show that such OOD sampling and pessimistic bootstrapping yields provable uncertainty quantifier in linear MDPs, thus providing the theoretical underpinning for PBRL. Extensive experiments on D4RL benchmark show that PBRL has better performance compared to the state-of-the-art algorithms.

Equivariant Subgraph Aggregation Networks

- Beatrice Bevilacqua, Fabrizio Frasca, Derek Lim, Balasubramaniam Srinivasan, Chen Cai, Gopinath Balamurugan, Michael M. Bronstein, Haggai Maron
- abstract@open-review(Spotlight): Message-passing neural networks (MPNNs) are the leading architecture for deep learning on graph-structured data, in large part due to their simplicity and scalability. Unfortunately, it was shown that these architectures are limited in their expressive power. This paper proposes a novel framework called Equivariant Subgraph Aggregation Networks (ESAN) to address this issue. Our main observation is that while two graphs may not be distinguishable by an MPNN, they often contain distinguishable subgraphs. Thus, we propose to represent each graph as a set of subgraphs derived by some predefined policy, and to process it using a suitable equivariant architecture. We develop novel variants of the 1-dimensional Weisfeiler-Leman (1-WL) test for graph isomorphism, and prove lower bounds on the expressiveness of ESAN in terms of these new WL variants. We further prove that our approach increases the expressive power of both MPNNs and more expressive architectures. Moreover, we provide theoretical results that describe how design choices such as the subgraph selection policy and equivariant neural architecture affect our architecture's expressive power. To deal with the increased computational cost, we propose a subgraph sampling scheme, which can be viewed as a stochastic version of our framework. A comprehensive set of experiments on real and synthetic datasets demonstrates that our framework improves the expressive power and overall performance of popular GNN architectures.

How Do Vision Transformers Work?

- Namuk Park, Songkuk Kim
- abstract@open-review(Spotlight): The success of multi-head self-attentions (MSAs) for computer vision is now indisputable. However, little is known about how MSAs work. We present fundamental explanations to help better understand the nature of MSAs. In particular, we demonstrate the following properties of MSAs and Vision Transformers (ViTs): (1) MSAs improve not only accuracy but also generalization by flattening the loss landscapes. Such

improvement is primarily attributable to their data specificity, not long-range dependency. On the other hand, ViTs suffer from non-convex losses. Large datasets and loss landscape smoothing methods alleviate this problem; (2) MSAs and Convs exhibit opposite behaviors. For example, MSAs are low-pass filters, but Convs are high-pass filters. Therefore, MSAs and Convs are complementary; (3) Multi-stage neural networks behave like a series connection of small individual models. In addition, MSAs at the end of a stage play a key role in prediction. Based on these insights, we propose AlterNet, a model in which Conv blocks at the end of a stage are replaced with MSA blocks. AlterNet outperforms CNNs not only in large data regimes but also in small data regimes. The code is available at https://github.com/xxxnell/how-do-vits-work.

Variational methods for simulation-based inference

- Manuel GlÃ¶ckler, Michael Deistler, Jakob H. Macke
- abstract@open-review(Spotlight): We present Sequential Neural Variational Inference (SNVI), an approach to perform Bayesian inference in models with intractable likelihoods. SNVI combines likelihood-estimation (or likelihood-ratio-estimation) with variational inference to achieve a scalable simulation-based inference approach. SNVI maintains the flexibility of likelihood(-ratio) estimation to allow arbitrary proposals for simulations, while simultaneously providing a functional estimate of the posterior distribution without requiring MCMC sampling. We present several variants of SNVI and demonstrate that they are substantially more computationally efficient than previous algorithms, without loss of accuracy on benchmark tasks. We apply SNVI to a neuroscience model of the pyloric network in the crab and demonstrate that it can infer the posterior distribution with one order of magnitude fewer simulations than previously reported. SNVI vastly reduces the computational cost of simulation-based inference while maintaining accuracy and flexibility, making it possible to tackle problems that were previously inaccessible.

Tackling the Generative Learning Trilemma with Denoising Diffusion GANs

- Zhisheng Xiao, Karsten Kreis, Arash Vahdat
- abstract@open-review(Spotlight): A wide variety of deep generative models has been developed in the past decade. Yet, these models often struggle with simultaneously addressing three key requirements including: high sample quality, mode coverage, and fast sampling. We call the challenge imposed by these requirements the generative learning trilemma, as the existing models often trade some of them for others. Particularly, denoising diffusion models have shown impressive sample quality and diversity, but their expensive sampling does not yet allow them to be applied in many real-world applications. In this paper, we argue that slow sampling in these models is fundamentally attributed to the Gaussian assumption in the denoising step which is justified only for small step sizes. To enable denoising with large steps, and hence, to reduce the total number of denoising steps, we propose to model the denoising distribution using a complex multimodal distribution. We introduce denoising diffusion generative adversarial networks (denoising diffusion GANs) that model each denoising step using a multimodal conditional GAN. Through extensive evaluations, we show that denoising diffusion GANs obtain sample quality and diversity competitive with original diffusion models while being 2000$\times$ faster on the CIFAR-10 dataset. Compared to traditional GANs, our model exhibits better mode coverage and sample diversity. To the best of our knowledge, denoising diffusion GAN is the first model that reduces sampling cost in diffusion models to an extent that allows them to be applied to real-world applications inexpensively.

Imbedding Deep Neural Networks

- Andrew Corbett, Dmitry Kangin
- abstract@open-review(Spotlight): Continuous-depth neural networks, such as Neural ODEs, have refashioned the understanding of residual neural networks in terms of non-linear vector-valued optimal control problems. The common solution is to use the adjoint sensitivity method to replicate a forward-backward pass optimisation problem. We propose a new approach which explicates the network's `depth' as a fundamental variable, thus reducing the problem to a system of forward-facing initial value problems. This new method is based on the principal of'Invariant Imbedding' for which we prove a general solution, applicable to all non-linear, vector-valued optimal control problems with both running and terminal loss. Our new architectures provide a tangible tool for inspecting the theoretical--and to a great extent unexplained--properties of network depth. They also constitute a resource of discrete implementations of Neural ODEs comparable to classes of imbedded residual neural networks. Through a series of experiments, we show the competitive performance of the proposed architectures for supervised learning and time series prediction.

Understanding and Preventing Capacity Loss in Reinforcement Learning

- Clare Lyle, Mark Rowland, Will Dabney
- abstract@open-review(Spotlight): The reinforcement learning (RL) problem is rife with sources of non-stationarity that can destabilize or inhibit learning progress. We identify a key mechanism by which this occurs in agents using neural networks as function approximators: \textit{capacity loss}, whereby networks trained to predict a sequence of target values lose their ability to quickly fit new functions over time. We demonstrate that capacity loss occurs in a broad range of RL agents and environments, and is particularly damaging to learning progress in sparse-reward tasks. We then present a simple regularizer, Initial Feature Regularization (InFeR), that mitigates this phenomenon by regressing a subspace of features towards its value at initialization, improving performance over a state-of-the-art model-free algorithm in the Atari 2600 suite. Finally, we study how this regularization affects different notions of capacity and evaluate other mechanisms by which it may improve performance.

Source-Free Adaptation to Measurement Shift via Bottom-Up Feature Restoration

- Cian Eastwood, Ian Mason, Chris Williams, Bernhard SchÃ¶lkopf
- abstract@open-review(Spotlight): Source-free domain adaptation (SFDA) aims to adapt a model trained on labelled data in a source domain to unlabelled data in a target domain without access to the source-domain data during adaptation. Existing methods for SFDA leverage entropy-minimization techniques which: (i) apply only to classification; (ii) destroy model calibration; and (iii) rely on the source model achieving a good level of feature-space class-separation in the target domain. We address these issues for a particularly pervasive type of domain shift called measurement shift which can be resolved by restoring the source features rather than extracting new ones. In particular, we propose Feature Restoration (FR) wherein we: (i) store a lightweight and flexible approximation of the feature distribution under the source data; and (ii) adapt the feature-extractor such that the approximate feature distribution under the target data realigns with that saved on the source. We additionally propose a bottom-up training scheme which boosts performance, which we call Bottom-Up Feature Restoration (BUFR). On real and synthetic data, we demonstrate that BUFR outperforms existing SFDA methods in terms of accuracy, calibration, and data efficiency, while being less reliant on the performance of the source model in the target domain.

The Inductive Bias of In-Context Learning: Rethinking Pretraining Example Design

- Yoav Levine, Noam Wies, Daniel Jannai, Dan Navon, Yedid Hoshen, Amnon Shashua
- abstract@open-review(Spotlight): Pretraining Neural Language Models (NLMs) over a large corpus involves chunking the text into training examples, which are contiguous text segments of sizes processable by the neural architecture. We highlight a bias introduced by this common practice: we prove that the pretrained NLM can model much stronger dependencies between text segments that appeared in the same training example, than it can between text segments that appeared in different training examples. This intuitive result has a twofold role. First, it formalizes the motivation behind a broad line of recent successful NLM training heuristics, proposed for the pretraining and fine-tuning stages, which do not necessarily appear related at first glance. Second, our result clearly indicates further improvements to be made in NLM pretraining for the benefit of Natural Language Understanding tasks. As an example, we propose ``kNN-Pretraining": we show that including semantically related non-neighboring sentences in the same pretraining example yields improved sentence representations and open domain question answering abilities. This theoretically motivated degree of freedom for pretraining example design indicates new training schemes for self-improving representations.

Emergent Communication at Scale

- Rahma Chaabouni, Florian Strub, Florent AltchÃ©, Eugene Tarassov, Corentin Tallec, Elnaz Davoodi, Kory Wallace Mathewson, Olivier Tieleman, Angeliki Lazaridou, Bilal Piot
- abstract@open-review(Spotlight): Emergent communication aims for a better understanding of human language evolution and building more efficient representations. We posit that reaching these goals will require scaling up, in contrast to a significant amount of literature that focuses on setting up small-scale problems to tease out desired properties of the emergent languages. We focus on three independent aspects to scale up, namely the dataset, task complexity, and population size. We provide a first set of results for large populations solving complex tasks on realistic large-scale datasets, as well as an easy-to-use codebase to enable further experimentation. In more complex tasks and datasets, we find that RL training can become unstable, but responds well to established stabilization techniques. We also identify the need for a different metric than topographic similarity, which does not correlate with the generalization performances when working with natural images. In this context, we probe ease-of-learnability and transfer methods to assess emergent languages. Finally, we observe that larger populations do not induce robust emergent protocols with high generalization performance, leading us to explore different ways to leverage population, through voting and imitation learning.

Superclass-Conditional Gaussian Mixture Model For Learning Fine-Grained Embeddings

- Jingchao Ni, Wei Cheng, Zhengzhang Chen, Takayoshi Asakura, Tomoya Soma, Sho Kato, Haifeng Chen
- abstract@open-review(Spotlight): Learning fine-grained embeddings is essential for extending the generalizability of models pre-trained on "coarse" labels (e.g., animals). It is crucial to fields for which fine-grained labeling (e.g., breeds of animals) is expensive, but fine-grained prediction is desirable, such as medicine. The dilemma necessitates adaptation of a "coarsely" pre-trained model to new tasks with a few "finer-grained" training labels. However, coarsely supervised pre-training tends to suppress intra-class variation, which is vital for cross-granularity adaptation. In this paper, we develop a training framework underlain by a novel superclass-conditional Gaussian mixture model (SCGM). SCGM imitates the generative process of samples from hierarchies of classes through latent variable modeling of the fine-grained subclasses. The framework is agnostic to the encoders and only adds a few distribution related parameters, thus is efficient, and flexible to different domains. The model parameters are learned end-to-end by maximum-likelihood estimation via a principled Expectation-Maximization algorithm. Extensive experiments on benchmark datasets and a real-life medical dataset indicate the effectiveness of our method.

SHINE: SHaring the INverse Estimate from the forward pass for bi-level optimization and implicit models

- Zaccharie Ramzi, Florian Mannel, Shaojie Bai, Jean-Luc Starck, Philippe Ciuciu, Thomas Moreau
- abstract@open-review(Spotlight): In recent years, implicit deep learning has emerged as a method to increase the depth of deep neural networks. While their training is memory-efficient, they are still significantly slower to train than their explicit counterparts. In Deep Equilibrium Models~(DEQs), the training is performed as a bi-level problem, and its computational complexity is partially driven by the iterative inversion of a huge Jacobian matrix. In this paper, we propose a novel strategy to tackle this computational bottleneck from which many bi-level problems suffer. The main idea is to use the quasi-Newton matrices from the forward pass to efficiently approximate the inverse Jacobian matrix in the direction needed for the gradient computation. We provide a theorem that motivates using our method with the original forward algorithms. In addition, by modifying these forward algorithms, we further provide theoretical guarantees that our method asymptotically estimates the true implicit gradient. We empirically study this approach in many settings, ranging from hyperparameter optimization to large Multiscale DEQs applied to CIFAR and ImageNet. We show that it reduces the computational cost of the backward pass by up to two orders of magnitude. All this is achieved while retaining the excellent performance of the original models in hyperparameter optimization and on CIFAR, and giving encouraging and competitive results on ImageNet.

Towards Deployment-Efficient Reinforcement Learning: Lower Bound and Optimality

- Jiawei Huang, Jinglin Chen, Li Zhao, Tao Qin, Nan Jiang, Tie-Yan Liu
- abstract@open-review(Spotlight): Deployment efficiency is an important criterion for many real-world applications of reinforcement learning (RL). Despite the community's increasing interest, there lacks a formal theoretical formulation for the problem. In this paper, we propose such a formulation for deployment-efficient RL (DE-RL) from an "optimization with constraints" perspective: we are interested in exploring an MDP and obtaining a near-optimal policy within minimal \emph{deployment complexity}, whereas in each deployment the policy can sample a large batch of data. Using finite-horizon linear MDPs as a concrete structural model, we reveal the fundamental limit in achieving deployment efficiency by establishing information-theoretic lower bounds, and provide algorithms that achieve the optimal deployment efficiency. Moreover, our formulation for DE-RL is flexible and can serve as a building block for other practically relevant settings; we give "Safe DE-RL" and "Sample-Efficient DE-RL" as two examples, which may be worth future investigation.

Task Relatedness-Based Generalization Bounds for Meta Learning

- Jiechao Guan, Zhiwu Lu
- abstract@open-review(Spotlight): Supposing the n training tasks and the new task are sampled from the same environment, traditional meta learning theory derives an error bound on the expected loss over the new task in terms of the empirical training loss, uniformly over the set of all hypothesis spaces. However, there is still little research on how the relatedness of these tasks can affect the full utilization of all mn training data (with m examples per task). In this paper, we propose to address this problem by defining a new notion of task relatedness according to the existence of the bijective transformation between two tasks. A novel generalization bound of $\mathcal{O}(\frac{1}{\sqrt{mn}})$ for meta learning is thus derived by exploiting the proposed task relatedness. Moreover, when investigating a special branch of meta learning that involves representation learning with deep neural networks, we establish spectrally-normalized bounds for both classification and regression problems. Finally, we demonstrate that the relatedness requirement between two tasks is satisfied when the sample space possesses the completeness and separability properties, validating the rationality and applicability of our proposed task-relatedness measure.

IntSGD: Adaptive Floatless Compression of Stochastic Gradients

- Konstantin Mishchenko, Bokun Wang, Dmitry Kovalev, Peter RichtÃ¡rik
- abstract@open-review(Spotlight): We propose a family of adaptive integer compression operators for distributed Stochastic Gradient Descent (SGD) that do not communicate a single float. This is achieved by multiplying floating-point vectors with a number known to every device and then rounding to integers. In contrast to the prior work on integer compression for SwitchML by (Sapio et al., 2021), our IntSGD method is provably convergent and computationally cheaper as it estimates the scaling of vectors adaptively. Our theory shows that the iteration complexity of IntSGD matches that of SGD up to constant factors for both convex and non-convex, smooth and non-smooth functions, with and without overparameterization. Moreover, our algorithm can also be tailored for the popular all-reduce primitive and shows promising empirical performance.

PAC-Bayes Information Bottleneck

- Zifeng Wang, Shao-Lun Huang, Ercan Engin Kuruoglu, Jimeng Sun, Xi Chen, Yefeng Zheng
- abstract@open-review(Spotlight): Understanding the source of the superior generalization ability of NNs remains one of the most important problems in ML research. There have been a series of theoretical works trying to derive non-vacuous bounds for NNs. Recently, the compression of information stored

in weights (IIW) is proved to play a key role in NNs generalization based on the PAC-Bayes theorem. However, no solution of IIW has ever been provided, which builds a barrier for further investigation of the IIW's property and its potential in practical deep learning. In this paper, we propose an algorithm for the efficient approximation of IIW. Then, we build an IIW-based information bottleneck on the trade-off between accuracy and information complexity of NNs, namely PIB. From PIB, we can empirically identify the fitting to compressing phase transition during NNs' training and the concrete connection between the IIW compression and the generalization. Besides, we verify that IIW is able to explain NNs in broad cases, e.g., varying batch sizes, over-parameterization, and noisy labels. Moreover, we propose an MCMC-based algorithm to sample from the optimal weight posterior characterized by PIB, which fulfills the potential of IIW in enhancing NNs in practice.

Byzantine-Robust Learning on Heterogeneous Datasets via Bucketing

- Sai Praneeth Karimireddy, Lie He, Martin Jaggi
- abstract@open-review(Spotlight): In Byzantine robust distributed or federated learning, a central server wants to train a machine learning model over data distributed across multiple workers. However, a fraction of these workers may deviate from the prescribed algorithm and send arbitrary messages. While this problem has received significant attention recently, most current defenses assume that the workers have identical data. For realistic cases when the data across workers are heterogeneous (non-iid), we design new attacks which circumvent current defenses, leading to significant loss of performance. We then propose a simple bucketing scheme that adapts existing robust algorithms to heterogeneous datasets at a negligible computational cost. We also theoretically and experimentally validate our approach, showing that combining bucketing with existing robust algorithms is effective against challenging attacks. Our work is the first to establish guaranteed convergence for the non-iid Byzantine robust problem under realistic assumptions.

Understanding the Role of Self Attention for Efficient Speech Recognition

- Kyuhong Shim, Jungwook Choi, Wonyong Sung
- abstract@open-review(Spotlight): Self-attention (SA) is a critical component of Transformer neural networks that have succeeded in automatic speech recognition (ASR). In this paper, we analyze the role of SA in Transformer-based ASR models for not only understanding the mechanism of improved recognition accuracy but also lowering the computational complexity. We reveal that SA performs two distinct roles: phonetic and linguistic localization. Especially, we show by experiments that phonetic localization in the lower layers extracts phonologically meaningful features from speech and reduces the phonetic variance in the utterance for proper linguistic localization in the upper layers. From this understanding, we discover that attention maps can be reused as long as their localization capability is preserved. To evaluate this idea, we implement the layer-wise attention map reuse on real GPU platforms and achieve up to 1.96 times speedup in inference and 33% savings in training time with noticeably improved ASR performance for the challenging benchmark on LibriSpeech dev/test-other dataset.

Label Encoding for Regression Networks

- Deval Shah, Zi Yu Xue, Tor Aamodt
- abstract@open-review(Spotlight): Deep neural networks are used for a wide range of regression problems. However, there exists a significant gap in accuracy between specialized approaches and generic direct regression in which a network is trained by minimizing the squared or absolute error of output labels. Prior work has shown that solving a regression problem with a set of binary classifiers can improve accuracy by utilizing well-studied binary classification algorithms. We introduce binary-encoded labels (BEL), which generalizes the application of binary classification to regression by providing a framework for considering arbitrary multi-bit values when encoding target values. We identify desirable properties of suitable encoding and decoding functions used for the conversion between real-valued and binary-encoded labels based on theoretical and empirical study. These properties highlight a tradeoff between classification error probability and error-correction capabilities of label encodings. BEL can be combined with off-the-shelf task-specific feature extractors and trained end-to-end. We propose a series of sample encoding, decoding, and training loss functions for BEL and demonstrate they result in lower error than direct regression and specialized approaches while being suitable for a diverse set of regression problems, network architectures, and evaluation metrics. BEL achieves state-of-the-art accuracies for several regression benchmarks. Code is available at https://github.com/ubc-aamodt-group/BEL_regression.

Equivariant Transformers for Neural Network based Molecular Potentials

- Philipp ThÃ¶lke, Gianni De Fabritiis
- abstract@open-review(Spotlight): The prediction of quantum mechanical properties is historically plagued by a trade-off between accuracy and speed. Machine learning potentials have previously shown great success in this domain, reaching increasingly better accuracy while maintaining computational efficiency comparable with classical force fields. In this work we propose TorchMD-NET, a novel equivariant Transformer (ET) architecture, outperforming state-of-the-art on MD17, ANI-1, and many QM9 targets in both accuracy and computational efficiency. Through an extensive attention weight analysis, we gain valuable insights into the black box predictor and show differences in the learned representation of conformers versus conformations sampled from molecular dynamics or normal modes. Furthermore, we highlight the importance of datasets including off-equilibrium conformations for the evaluation of molecular potentials.

SGD Can Converge to Local Maxima

- Liu Ziyin, Botao Li, James B Simon, Masahito Ueda
- abstract@open-review(Spotlight): Previous works on stochastic gradient descent (SGD) often focus on its success. In this work, we construct worst-case optimization problems illustrating that, when not in the regimes that the previous works often assume, SGD can exhibit many strange and potentially undesirable behaviors. Specifically, we construct landscapes and data distributions such that (1) SGD converges to local maxima, (2) SGD escapes saddle points arbitrarily slowly, (3) SGD prefers sharp minima over flat ones, and (4) AMSGrad converges to local maxima. We also realize results in a minimal neural network-like example. Our results highlight the importance of simultaneously analyzing the minibatch sampling, discrete-time updates rules, and realistic landscapes to understand the role of SGD in deep learning.

Hybrid Local SGD for Federated Learning with Heterogeneous Communications

- Yuanxiong Guo, Ying Sun, Rui Hu, Yanmin Gong
- abstract@open-review(Spotlight): Communication is a key bottleneck in federated learning where a large number of edge devices collaboratively learn a model under the orchestration of a central server without sharing their own training data. While local SGD has been proposed to reduce the number of FL rounds and become the algorithm of choice for FL, its total communication cost is still prohibitive when each device needs to communicate with the remote server repeatedly for many times over bandwidth-limited networks. In light of both device-to-device (D2D) and device-to-server (D2S) cooperation opportunities in modern communication networks, this paper proposes a new federated optimization algorithm dubbed hybrid local SGD (HL-SGD) in FL settings where devices are grouped into a set of disjoint clusters with high D2D communication bandwidth. HL-SGD subsumes previous proposed algorithms such as local SGD and gossip SGD and enables us to strike the best balance between model accuracy and runtime. We analyze the convergence of HL-SGD in the presence of heterogeneous data for general nonconvex settings. We also perform extensive experiments and show that the use of hybrid model aggregation via D2D and D2S communications in HL-SGD can largely speed up the training time of federated learning.

Increasing the Cost of Model Extraction with Calibrated Proof of Work

- Adam Dziedzic, Muhammad Ahmad Kaleem, Yu Shen Lu, Nicolas Papernot
- abstract@open-review(Spotlight): In model extraction attacks, adversaries can steal a machine learning model exposed via a public API by repeatedly querying it and adjusting their own model based on obtained predictions. To prevent model stealing, existing defenses focus on detecting malicious queries, truncating, or distorting outputs, thus necessarily introducing a tradeoff between robustness and model utility for legitimate users. Instead, we propose to impede model extraction by requiring users to complete a proof-of-work before they can read the model's predictions. This deters attackers by greatly increasing (even up to 100x) the computational effort needed to leverage query access for model extraction. Since we calibrate the effort required to complete the proof-of-work to each query, this only introduces a slight overhead for regular users (up to 2x). To achieve this, our calibration applies tools from differential privacy to measure the information revealed by a query. Our method requires no modification of the victim model and can be applied by machine learning practitioners to guard their publicly exposed models against being easily stolen.

Learning the Dynamics of Physical Systems from Sparse Observations with Finite Element Networks

- Marten Lienen, Stephan GÃ¼nnemann
- abstract@open-review(Spotlight): We propose a new method for spatio-temporal forecasting on arbitrarily distributed points. Assuming that the observed system follows an unknown partial differential equation, we derive a continuous-time model for the dynamics of the data via the finite element method. The resulting graph neural network estimates the instantaneous effects of the unknown dynamics on each cell in a meshing of the spatial domain. Our model can incorporate prior knowledge via assumptions on the form of the unknown PDE, which induce a structural bias towards learning specific processes. Through this mechanism, we derive a transport variant of our model from the convection equation and show that it improves the transfer performance to higher-resolution meshes on sea surface temperature and gas flow forecasting against baseline models representing a selection of spatio-temporal forecasting methods. A qualitative analysis shows that our model disentangles the data dynamics into their constituent parts, which makes it uniquely interpretable.

Probabilistic Implicit Scene Completion

- Dongsu Zhang, Changwoon Choi, Inbum Park, Young Min Kim
- abstract@open-review(Spotlight): We propose a probabilistic shape completion method extended to the continuous geometry of large-scale 3D scenes. Real-world scans of 3D scenes suffer from a considerable amount of missing data cluttered with unsegmented objects. The problem of shape completion is inherently ill-posed, and high-quality result requires scalable solutions that consider multiple possible outcomes. We employ the Generative Cellular Automata that learns the multi-modal distribution and transform the formulation to process large-scale continuous geometry. The local continuous shape is incrementally generated as a sparse voxel embedding, which contains the latent code for each occupied cell. We formally derive that our training objective for the sparse voxel embedding maximizes the variational lower bound of the complete shape distribution and therefore our progressive generation constitutes a valid generative model. Experiments show that our model successfully generates diverse plausible scenes faithful to the input, especially when the input suffers from a significant amount of missing data. We also demonstrate that our approach outperforms deterministic models even in less ambiguous cases with a small amount of missing data, which infers that probabilistic formulation is crucial for high-quality geometry completion on input scans exhibiting any levels of completeness.

Improving Federated Learning Face Recognition via Privacy-Agnostic Clusters

- Qiang Meng, Feng Zhou, Hainan Ren, Tianshu Feng, Guochao Liu, Yuanqing Lin
- abstract@open-review(Spotlight): The growing public concerns on data privacy in face recognition can be partly relieved by the federated learning (FL) paradigm. However, conventional FL methods usually perform poorly due to the particularity of the task, \textit{i.e.}, broadcasting class centers among clients is essential for recognition performances but leads to privacy leakage. To resolve the privacy-utility paradox, this work proposes PrivacyFace, a framework largely improves the federated learning face recognition via communicating auxiliary and privacy-agnostic information among clients. PrivacyFace mainly consists of two components: First, a practical Differentially Private Local Clustering (DPLC) mechanism is proposed to distill sanitized clusters from local class centers. Second, a consensus-aware recognition loss subsequently encourages global consensuses among clients, which ergo leads to more discriminative features. The proposed schemes are mathematically proved to be differential private, introduce a lightweight overhead as well as yield prominent performance boosts (\textit{e.g.}, +9.63\% and +10.26\% for TAR@FAR=1e-4 on IJB-B and IJB-C respectively). Extensive experiments and ablation studies on a large-scale dataset have demonstrated the efficacy and practicability of our method.

Learning to Downsample for Segmentation of Ultra-High Resolution Images

- Chen Jin, Ryutaro Tanno, Thomy Mertzanidou, Eleftheria Panagiotaki, Daniel C. Alexander
- abstract@open-review(Poster): Many computer vision systems require low-cost segmentation algorithms based on deep learning, either because of the enormous size of input images or limited computational budget. Common solutions uniformly downsample the input images to meet memory constraints, assuming all pixels are equally informative. In this work, we demonstrate that this assumption can harm the segmentation performance because the segmentation difficulty varies spatially (see Figure 1 â€œUniformâ€). We combat this problem by introducing a learnable downsampling module, which can be optimised together with the given segmentation model in an end-to-end fashion. We formulate the problem of training such downsampling module as optimisation of sampling density distributions over the input images given their low-resolution views. To defend against degenerate solutions (e.g. over-sampling trivial regions like the backgrounds), we propose a regularisation term that encourages the sampling locations to concentrate around the object boundaries. We find the downsampling module learns to sample more densely at difficult locations, thereby improving the segmentation performance (see Figure 1 "Ours"). Our experiments on benchmarks of high-resolution street view, aerial and medical images demonstrate substantial improvements in terms of efficiency-and-accuracy trade-off compared to both uniform downsampling and two recent advanced downsampling techniques.

Variational Neural Cellular Automata

- Rasmus Berg Palm, Miguel GonzÃ¡lez Duque, Shyam Sudhakaran, Sebastian Risi
- abstract@open-review(Poster): In nature, the process of cellular growth and differentiation has lead to an amazing diversity of organisms --- algae, starfish, giant sequoia, tardigrades, and orcas are all created by the same generative process. Inspired by the incredible diversity of this biological generative process, we propose a generative model, the Variational Neural Cellular Automata (VNCA), which is loosely inspired by the biological processes of cellular growth and differentiation. Unlike previous related works, the VNCA is a proper probabilistic generative model, and we evaluate it according to best practices. We find that the VNCA learns to reconstruct samples well and that despite its relatively few parameters and simple local-only communication, the VNCA can learn to generate a large variety of output from information encoded in a common vector format. While there is a significant gap to the current state-of-the-art in terms of generative modeling performance, we show that the VNCA can learn a purely self-organizing generative process of data. Additionally, the self-organizing nature bestows the VNCA with some inherent robustness against perturbations in the early stages of growth.

Wish you were here: Hindsight Goal Selection for long-horizon dexterous manipulation

- Todor Davchev, Oleg Olegovich Sushkov, Jean-Baptiste Regli, Stefan Schaal, Yusuf Aytar, Markus Wulfmeier, Jon Scholz
- abstract@open-review(Poster): Complex sequential tasks in continuous-control settings often require agents to successfully traverse a set of ``narrow passages'' in their state space. Solving such tasks with a sparse reward in a sample-efficient manner poses a challenge to modern reinforcement learning (RL) due to the associated long-horizon nature of the problem and the lack of sufficient positive signal during learning. Various tools have been applied to

address this challenge. When available, large sets of demonstrations can guide agent exploration. Hindsight relabelling on the other hand does not require additional sources of information. However, existing strategies explore based on task-agnostic goal distributions, which can render the solution of long-horizon tasks impractical. In this work, we extend hindsight relabelling mechanisms to guide exploration along task-specific distributions implied by a small set of successful demonstrations. We evaluate the approach on four complex, single and dual arm, robotics manipulation tasks against strong suitable baselines. The method requires far fewer demonstrations to solve all tasks and achieves a significantly higher overall performance as task complexity increases. Finally, we investigate the robustness of the proposed solution with respect to the quality of input representations and the number of demonstrations.

L0-Sparse Canonical Correlation Analysis

- Ofir Lindenbaum, Moshe Salhov, Amir Averbuch, Yuval Kluger
- abstract@open-review(Poster): Canonical Correlation Analysis (CCA) models are powerful for studying the associations between two sets of variables. The canonically correlated representations, termed \textit{canonical variates} are widely used in unsupervised learning to analyze unlabeled multi-modal registered datasets. Despite their success, CCA models may break (or overfit) if the number of variables in either of the modalities exceeds the number of samples. Moreover, often a significant fraction of the variables measures modality-specific information, and thus removing them is beneficial for identifying the \textit{canonically correlated variates}. Here, we propose ℓ_0-CCA, a method for learning correlated representations based on sparse subsets of variables from two observed modalities. Sparsity is obtained by multiplying the input variables by stochastic gates, whose parameters are learned together with the CCA weights via an ℓ_0-regularized correlation loss. We further propose ℓ_0-Deep CCA for solving the problem of non-linear sparse CCA by modeling the correlated representations using deep nets. We demonstrate the efficacy of the method using several synthetic and real examples. Most notably, by gating nuisance input variables, our approach improves the extracted representations compared to other linear, non-linear and sparse CCA-based models.

Recycling Model Updates in Federated Learning: Are Gradient Subspaces Low-Rank?

- Sheikh Shams Azam, Seyyedali Hosseinalipour, Qiang Qiu, Christopher Brinton
- abstract@open-review(Poster): In this paper, we question the rationale behind propagating large numbers of parameters through a distributed system during federated learning. We start by examining the rank characteristics of the subspace spanned by gradients (i.e., the gradient-space) in centralized model training, and observe that the gradient-space often consists of a few leading principal components accounting for an overwhelming majority (95-99%) of the explained variance. Motivated by this, we propose the "Look-back Gradient Multiplier" (LBGM) algorithm, which utilizes this low-rank property of the gradient-space in federated learning. Operationally, LBGM recycles the gradients between model update rounds to significantly reduce the number of parameters to be propagated through the system. We analytically characterize the convergence behavior of LBGM, revealing the nature of the trade-off between communication savings and model performance. Our subsequent experimental results demonstrate the improvement LBGM obtains on communication overhead compared to federated learning baselines. Additionally, we show that LBGM is a general plug-and-play algorithm that can be used standalone or stacked on top of existing sparsification techniques for distributed model training.

Is Homophily a Necessity for Graph Neural Networks?

- Yao Ma, Xiaorui Liu, Neil Shah, Jiliang Tang
- abstract@open-review(Poster): Graph neural networks (GNNs) have shown great prowess in learning representations suitable for numerous graph-based machine learning tasks. When applied to semi-supervised node classification, GNNs are widely believed to work well due to the homophily assumption (``like attracts like"), and fail to generalize to heterophilous graphs where dissimilar nodes connect. Recent works design new architectures to overcome such heterophily-related limitations, citing poor baseline performance and new architecture improvements on a few heterophilous graph benchmark datasets as evidence for this notion. In our experiments, we empirically find that standard graph convolutional networks (GCNs) can actually achieve better performance than such carefully designed methods on some commonly used heterophilous graphs. This motivates us to reconsider whether homophily is truly necessary for good GNN performance. We find that this claim is not quite true, and in fact, GCNs can achieve strong performance on heterophilous graphs under certain conditions. Our work carefully characterizes these conditions and provides supporting theoretical understanding and empirical observations. Finally, we examine existing heterophilous graphs benchmarks and reconcile how the GCN (under)performs on them based on this understanding.

DEGREE: Decomposition Based Explanation for Graph Neural Networks

- Qizhang Feng, Ninghao Liu, Fan Yang, Ruixiang Tang, Mengnan Du, Xia Hu
- abstract@open-review(Poster): Graph Neural Networks (GNNs) are gaining extensive attention for their application in graph data. However, the black-box nature of GNNs prevents users from understanding and trusting the models, thus hampering their applicability. Whereas explaining GNNs remains a challenge, most existing methods fall into approximation based and perturbation based approaches with suffer from faithfulness problems and unnatural artifacts respectively. To tackle these problems, we propose DEGREE (Decomposition based Explanation for GRaph nEural nEtworks) to provide a faithful explanation for GNN predictions. By decomposing the information generation and aggregation mechanism of GNNs, DEGREE allows tracking the contributions of specific components of the input graph to the final prediction. Based on this, we further design a subgraph level interpretation algorithm to reveal complex interactions between graph nodes that are overlooked by previous methods. The efficiency of our algorithm can be further improved by utilizing GNN characteristics. Finally, we conduct quantitative and qualitative experiments on synthetic and real-world datasets to demonstrate the effectiveness of DEGREE on node classification and graph classification tasks.

Improving Mutual Information Estimation with Annealed and Energy-Based Bounds

- Rob Brekelmans, Sicong Huang, Marzyeh Ghassemi, Greg Ver Steeg, Roger Baker Grosse, Alireza Makhzani
- abstract@open-review(Poster): Mutual information (MI) is a fundamental quantity in information theory and machine learning. However, direct estimation of MI is intractable, even if the true joint probability density for the variables of interest is known, as it involves estimating a potentially high-dimensional log partition function. In this work, we present a unifying view of existing MI bounds from the perspective of importance sampling, and propose three novel bounds based on this approach. Since a tight MI bound without density information requires a sample size exponential in the true MI, we assume either a single marginal or the full joint density information is known. In settings where the full joint density is available, we propose Multi-Sample Annealed Importance Sampling (AIS) bounds on MI, which we demonstrate can tightly estimate large values of MI in our experiments. In settings where only a single marginal distribution is known, we propose Generalized IWAE (GIWAE) and MINE-AIS bounds. Our GIWAE bound unifies variational and contrastive bounds in a single framework that generalizes InfoNCE, IWAE, and Barber-Agakov bounds. Our MINE-AIS method improves upon existing energy-based methods such as MINE-DV and MINE-F by directly optimizing a tighter lower bound on MI. MINE-AIS uses MCMC sampling to estimate gradients for training and Multi-Sample AIS for evaluating the bound. Our methods are particularly suitable for evaluating MI in deep generative models, since explicit forms of the marginal or joint densities are often available. We evaluate our bounds on estimating the MI of VAEs and GANs trained on the MNIST and CIFAR datasets, and showcase significant gains over existing bounds in these challenging settings with high ground truth MI.

Sequence Approximation using Feedforward Spiking Neural Network for Spatiotemporal Learning: Theory and Optimization Methods

- Xueyuan She, Saurabh Dash, Saibal Mukhopadhyay
- abstract@open-review(Poster): A dynamical system of spiking neurons with only feedforward connections can classify spatiotemporal patterns without recurrent connections. However, the theoretical construct of a feedforward spiking neural network (SNN) for approximating a temporal sequence remains unclear, making it challenging to optimize SNN architectures for learning complex spatiotemporal patterns. In this work, we establish a theoretical framework to understand and improve sequence approximation using a feedforward SNN. Our framework shows that a feedforward SNN with one neuron per layer and skip-layer connections can approximate the mapping function between any arbitrary pairs of input and output spike train on a compact domain. Moreover, we prove that heterogeneous neurons with varying dynamics and skip-layer connections improve sequence approximation using feedforward SNN. Consequently, we propose SNN architectures incorporating the preceding constructs that are trained using supervised backpropagation-through-time (BPTT) and unsupervised spiking-timing-dependent plasticity (STDP) algorithms for classification of spatiotemporal data. A dual-search-space Bayesian optimization method is developed to optimize architecture and parameters of the proposed SNN with heterogeneous neuron dynamics and skip-layer connections.

Diverse Client Selection for Federated Learning via Submodular Maximization

- Ravikumar Balakrishnan, Tian Li, Tianyi Zhou, Nageen Himayat, Virginia Smith, Jeff Bilmes
- abstract@open-review(Poster): In every communication round of federated learning, a random subset of clients communicate their model updates back to the server which then aggregates them all. The optimal size of this subset is not known and several studies have shown that typically random selection does not perform very well in terms of convergence, learning efficiency and fairness. We, in this paper, propose to select a small diverse subset of clients, namely those carrying representative gradient information, and we transmit only these updates to the server. Our aim is for updating via only a subset to approximate updating via aggregating all client information. We achieve this by choosing a subset that maximizes a submodular facility location function defined over gradient space. We introduce â€œfederated averaging with diverse client selection (DivFL)â€ . We provide a thorough analysis of its convergence in the heterogeneous setting and apply it both to synthetic and to real datasets. Empirical results show several benefits to our approach including improved learning efficiency, faster convergence and also more uniform (i.e., fair) performance across clients. We further show a communication-efficient version of DivFL that can still outperform baselines on the above metrics.

From Intervention to Domain Transportation: A Novel Perspective to Optimize Recommendation

- Da Xu, Yuting Ye, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, Kannan Achan
- abstract@open-review(Poster): The interventional nature of recommendation has attracted increasing attention in recent years. It particularly motivates researchers to formulate learning and evaluating recommendation as causal inference and data missing-not-at-random problems. However, few take seriously the consequence of violating the critical assumption of overlapping, which we prove can significantly threaten the validity and interpretation of the outcome. We find a critical piece missing in the current understanding of information retrieval (IR) systems: as interventions, recommendation not only affects the already observed data, but it also interferes with the target domain (distribution) of interest. We then rephrase optimizing recommendation as finding an intervention that best transports the patterns it learns from the observed domain to its intervention domain. Towards this end, we use domain transportation to characterize the learning-intervention mechanism of recommendation. We design a principled transportation-constraint risk minimization objective and convert it to a two-player minimax game. We prove the consistency, generalization, and excessive risk bounds for the proposed objective, and elaborate how they compare to the current results. Finally, we carry out extensive real-data and semi-synthetic experiments to demonstrate the advantage of our approach, and launch online testing with a real-world IR system.

Variational Predictive Routing with Nested Subjective Timescales

- Alexey Zakharov, Qinghai Guo, Zafeirios Fountas
- abstract@open-review(Poster): Discovery and learning of an underlying spatiotemporal hierarchy in sequential data is an important topic for machine learning. Despite this, little work has been done to explore hierarchical generative models that can flexibly adapt their layerwise representations in response to datasets with different temporal dynamics. Here, we present Variational Predictive Routing (VPR) â€“ a neural probabilistic inference system that organizes latent representations of video features in a temporal hierarchy, based on their rates of change, thus modeling continuous data as a hierarchical renewal process. By employing an event detection mechanism that relies solely on the systemâ€™s latent representations (without the need of a separate model), VPR is able to dynamically adjust its internal state following changes in the observed features, promoting an optimal organisation of representations across the levels of the modelâ€™s latent hierarchy. Using several video datasets, we show that VPR is able to detect event boundaries, disentangle spatiotemporal features across its hierarchy, adapt to the dynamics of the data, and produce accurate time-agnostic rollouts of the future. Our approach integrates insights from neuroscience and introduces a framework with high potential for applications in model-based reinforcement learning, where flexible and informative state-space rollouts are of particular interest.

Sample and Computation Redistribution for Efficient Face Detection

- Jia Guo, Jiankang Deng, Alexandros Lattas, Stefanos Zafeiriou
- abstract@open-review(Poster): Although tremendous strides have been made in uncontrolled face detection, accurate face detection with a low computation cost remains an open challenge. In this paper, we point out that computation distribution and scale augmentation are the keys to detecting small faces from low-resolution images. Motivated by these observations, we introduce two simple but effective methods: (1) Computation Redistribution (CR), which reallocates the computation between the backbone, neck and head of the model; and (2) Sample Redistribution (SR), which augments training samples for the most needed stages. The proposed Sample and Computation Redistribution for Face Detection (SCRFD) is implemented by a random search in a meticulously designed search space. Extensive experiments conducted on WIDER FACE demonstrate the state-of-the-art accuracy-efficiency trade-off for the proposed SCRFD family across a wide range of compute regimes. In particular, SCRFD-34GF outperforms the best competitor, TinaFace, by 4.78% (AP at hard set) while being more than 3$\times$ faster on GPUs with VGA-resolution images. Code is available at: https://github.com/deepinsight/insightface/tree/master/detection/scrfd.

Sound Adversarial Audio-Visual Navigation

- Yinfeng Yu, Wenbing Huang, Fuchun Sun, Changan Chen, Yikai Wang, Xiaohong Liu
- abstract@open-review(Poster): Audio-visual navigation task requires an agent to find a sound source in a realistic, unmapped 3D environment by utilizing egocentric audio-visual observations. Existing audio-visual navigation works assume a clean environment that solely contains the target sound, which, however, would not be suitable in most real-world applications due to the unexpected sound noise or intentional interference. In this work, we design an acoustically complex environment in which, besides the target sound, there exists a sound attacker playing a zero-sum game with the agent. More specifically, the attacker can move and change the volume and category of the sound to make the agent suffer from finding the sounding object while the agent tries to dodge the attack and navigate to the goal under the intervention. Under certain constraints to the attacker, we can improve the robustness of the agent towards unexpected sound attacks in audio-visual navigation. For better convergence, we develop a joint training mechanism by employing the property of a centralized critic with decentralized actors. Experiments on two real-world 3D scan datasets, Replica, and Matterport3D, verify the effectiveness and the robustness of the agent trained under our designed environment when transferred to the clean environment or the one containing sound attackers with random policy. Project: https://yyf17.github.io/SAAVN .

Out-of-distribution Generalization in the Presence of Nuisance-Induced Spurious Correlations

- Aahlad Manas Puli, Lily H Zhang, Eric Karl Oermann, Rajesh Ranganath
- abstract@open-review(Poster): In many prediction problems, spurious correlations are induced by a changing relationship between the label and a nuisance variable that is also correlated with the covariates. For example, in classifying animals in natural images, the background, which is a nuisance, can predict the type of animal. This nuisance-label relationship does not always hold, and the performance of a model trained under one such relationship may be poor on data with a different nuisance-label relationship. To build predictive models that perform well regardless of the nuisance-label relationship, we develop Nuisance-Randomized Distillation (NURD). We introduce the nuisance-randomized distribution, a distribution where the nuisance and the label are independent. Under this distribution, we define the set of representations such that conditioning on any member, the nuisance and the label remain independent. We prove that the representations in this set always perform better than chance, while representations outside of this set may not. NURD finds a representation from this set that is most informative of the label under the nuisance-randomized distribution, and we prove that this representation achieves the highest performance regardless of the nuisance-label relationship. We evaluate NURD on several tasks including chest X-ray classification where, using non-lung patches as the nuisance, NURD produces models that predict pneumonia under strong spurious correlations.

AEVA: Black-box Backdoor Detection Using Adversarial Extreme Value Analysis

- Junfeng Guo, Ang Li, Cong Liu
- abstract@open-review(Poster): Deep neural networks (DNNs) are proved to be vulnerable against backdoor attacks. A backdoor could be embedded in the target DNNs through injecting a backdoor trigger into the training examples, which can cause the target DNNs misclassify an input attached with the backdoor trigger. Recent backdoor detection methods often require the access to the original poisoned training data, the parameters of the target DNNs, or the predictive confidence for each given input, which are impractical in many real-world applications, e.g., on-device de-ployed DNNs. We address the black-box hard-label backdoor detection problem where the DNN is a fully black-box and only its final output label is accessible. We approach this problem from the optimization perspective and show that the objective of backdoor detection is bounded by an adversarial objective. Further theoretical and empirical studies reveal that this adversarial objective leads to a solution with highly skewed distribution; a singularity is often observed in the adversarial map of a backdoor-infected example, which we call the adversarial singularity phenomenon. Based on this observation, we propose the adversarial extreme value analysis(AEVA) algorithm to detect backdoors in black-box neural networks. The AEVA algorithm is based on an extreme value analysis on the adversarial map, computed from the monte-carlo gradient estimation due to the black-box hard-label constraint. Evidenced by extensive experiments across three popular tasks and backdoor attacks, our approach is shown effective in detecting backdoor attacks under the black-box hard-label scenarios

Resonance in Weight Space: Covariate Shift Can Drive Divergence of SGD with Momentum

- Kirby Banman, Garnet Liam Peet-Pare, Nidhi Hegde, Alona Fyshe, Martha White
- abstract@open-review(Poster): Most convergence guarantees for stochastic gradient descent with momentum (SGDm) rely on iid sampling. Yet, SGDm is often used outside this regime, in settings with temporally correlated input samples such as continual learning and reinforcement learning. Existing work has shown that SGDm with a decaying step-size can converge under Markovian temporal correlation. In this work, we show that SGDm under covariate shift with a fixed step-size can be unstable and diverge. In particular, we show SGDm under covariate shift is a parametric oscillator, and so can suffer from a phenomenon known as resonance. We approximate the learning system as a time varying system of ordinary differential equations, and leverage existing theory to characterize the system's divergence/convergence as resonant/nonresonant modes. The theoretical result is limited to the linear setting with periodic covariate shift, so we empirically supplement this result to show that resonance phenomena persist even under non-periodic covariate shift, nonlinear dynamics with neural networks, and optimizers other than SGDm.

Top-label calibration and multiclass-to-binary reductions

- Chirag Gupta, Aaditya Ramdas
- abstract@open-review(Poster): We propose a new notion of multiclass calibration called top-label calibration. A classifier is said to be top-label calibrated if the reported probability for the predicted class label---the top-label---is calibrated, conditioned on the top-label. This conditioning is essential for practical utility of the calibration property, since the top-label is always reported and we must condition on what is reported. However, the popular notion of confidence calibration erroneously skips this conditioning. Furthermore, we outline a multiclass-to-binary (M2B) reduction framework that unifies confidence, top-label, and class-wise calibration, among others. As its name suggests, M2B works by reducing multiclass calibration to different binary calibration problems; various types of multiclass calibration can then be achieved using simple binary calibration routines. We instantiate the M2B framework with the well-studied histogram binning (HB) binary calibrator, and prove that the overall procedure is multiclass calibrated without making any assumptions on the underlying data distribution. In an empirical evaluation with four deep net architectures on CIFAR-10 and CIFAR-100, we find that the M2B + HB procedure achieves lower top-label and class-wise calibration error than other approaches such as temperature scaling. Code for this work is available at https://github.com/aigen/df-posthoc-calibration.

Anisotropic Random Feature Regression in High Dimensions

- Gabriel Mel, Jeffrey Pennington
- abstract@open-review(Poster): In contrast to standard statistical wisdom, modern learning algorithms typically find their best performance in the overparameterized regime in which the model has many more parameters than needed to fit the training data. A growing number of recent works have shown that random feature models can offer a detailed theoretical explanation for this unexpected behavior, but typically these analyses have utilized isotropic distributional assumptions on the underlying data generation process, thereby failing to provide a realistic characterization of real-world models that are designed to identify and harness the structure in natural data. In this work, we examine the high-dimensional asymptotics of random feature regression in the presence of structured data, allowing for arbitrary input correlations and arbitrary alignment between the data and the weights of the target function. We define a partial order on the space of weight-data alignments and prove that generalization performance improves in response to stronger alignment. We also clarify several previous observations in the literature by distinguishing the behavior of the sample-wise and parameter-wise learning curves, finding that sample-wise multiple descent can occur at scales dictated by the eigenstructure of the data covariance, but that parameter-wise multiple descent is limited to double descent, although strong anisotropy can induce additional signatures such as wide plateaus and steep cliffs. Finally, these signatures are related to phase transitions in the spectrum of the feature kernel matrix, and unlike the double descent peak, persist even under optimal regularization.

Back2Future: Leveraging Backfill Dynamics for Improving Real-time Predictions in Future

- Harshavardhan Kamarthi, Alexander RodrÃguez, B. Aditya Prakash
- abstract@open-review(Poster): For real-time forecasting in domains like public health and macroeconomics, data collection is a non-trivial and demanding task. Often after being initially released, it undergoes several revisions later (maybe due to human or technical constraints) - as a result, it may take weeks until the data reaches a stable value. This so-called â€˜backfillâ€™ phenomenon and its effect on model performance have been barely addressed in the prior literature. In this paper, we introduce the multi-variate backfill problem using COVID-19 as the motivating example. We construct a detailed dataset composed of relevant signals over the past year of the pandemic. We then systematically characterize several patterns in backfill dynamics and leverage our observations for formulating a novel problem and neural framework, Back2Future, that aims to refines a given model's predictions in real-time. Our extensive experiments demonstrate that our method refines the performance of the diverse set of top models for COVID-19 forecasting and GDP growth forecasting. Specifically, we show that Back2Future refined top COVID-19 models by 6.65% to 11.24% and yield an 18% improvement over

non-trivial baselines. In addition, we show that our model improves model evaluation too; hence policy-makers can better understand the true accuracy of forecasting models in real-time.

Approximation and Learning with Deep Convolutional Models: a Kernel Perspective

- Alberto Bietti
- abstract@open-review(Poster): The empirical success of deep convolutional networks on tasks involving high-dimensional data such as images or audio suggests that they can efficiently approximate certain functions that are well-suited for such tasks. In this paper, we study this through the lens of kernel methods, by considering simple hierarchical kernels with two or three convolution and pooling layers, inspired by convolutional kernel networks. These achieve good empirical performance on standard vision datasets, while providing a precise description of their functional space that yields new insights on their inductive bias. We show that the RKHS consists of additive models of interaction terms between patches, and that its norm encourages spatial similarities between these terms through pooling layers. We then provide generalization bounds which illustrate how pooling and patches yield improved sample complexity guarantees when the target function presents such regularities.

Value Function Spaces: Skill-Centric State Abstractions for Long-Horizon Reasoning

- Dhruv Shah, Peng Xu, Yao Lu, Ted Xiao, Alexander T Toshev, Sergey Levine, brian ichter
- abstract@open-review(Poster): Reinforcement learning can train policies that effectively perform complex tasks. However for long-horizon tasks, the performance of these methods degrades with horizon, often necessitating reasoning over and chaining lower-level skills. Hierarchical reinforcement learning aims to enable this by providing a bank of low-level skills as action abstractions. Hierarchies can further improve on this by abstracting the space states as well. We posit that a suitable state abstraction should depend on the capabilities of the available lower-level policies. We propose Value Function Spaces: a simple approach that produces such a representation by using the value functions corresponding to each lower-level skill. These value functions capture the affordances of the scene, thus forming a representation that compactly abstracts task relevant information and robustly ignores distractors. Empirical evaluations for maze-solving and robotic manipulation tasks demonstrate that our approach improves long-horizon performance and enables better zero-shot generalization than alternative model-free and model-based methods.

Fast Regression for Structured Inputs

- Raphael A Meyer, Cameron N Musco, Christopher P Musco, David Woodruff, Samson Zhou
- abstract@open-review(Poster): We study the ℓ_p regression problem, which requires finding $\mathbf{x}\in\mathbb R^{d}$ that minimizes $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p$ for a matrix $\mathbf{A}\in\mathbb R^{n \times d}$ and response vector $\mathbf{b}\in\mathbb R^{n}$. There has been recent interest in developing subsampling methods for this problem that can outperform standard techniques when n is very large. However, all known subsampling approaches have run time that depends exponentially on p, typically, $d^{\mathcal{O}(p)}$, which can be prohibitively expensive.

We improve on this work by showing that for a large class of common \emph{structured matrices}, such as combinations of low-rank matrices, sparse matrices, and Vandermonde matrices, there are subsampling based methods for ℓ_p regression that depend polynomially on p. For example, we give an algorithm for ℓ_p regression on Vandermonde matrices that runs in time $\mathcal{O}(n\log^3 n+(dp^2)^{0.5+\omega}\cdot\text{polylog}\,n)$, where ω is the exponent of matrix multiplication. The polynomial dependence on p crucially allows our algorithms to extend naturally to efficient algorithms for ℓ_∞ regression, via approximation of ℓ_∞ by $\ell_{\mathcal{O}(\log n)}$. Of practical interest, we also develop a new subsampling algorithm for ℓ_p regression for arbitrary matrices, which is simpler than previous approaches for $p \ge 4$.

CrossBeam: Learning to Search in Bottom-Up Program Synthesis

- Kensen Shi, Hanjun Dai, Kevin Ellis, Charles Sutton
- abstract@open-review(Poster): Many approaches to program synthesis perform a search within an enormous space of programs to find one that satisfies a given specification. Prior works have used neural models to guide combinatorial search algorithms, but such approaches still explore a huge portion of the search space and quickly become intractable as the size of the desired program increases. To tame the search space blowup, we propose training a neural model to learn a hands-on search policy for bottom-up synthesis, instead of relying on a combinatorial search algorithm. Our approach, called CrossBeam, uses the neural model to choose how to combine previously-explored programs into new programs, taking into account the search history and partial program executions. Motivated by work in structured prediction on learning to search, CrossBeam is trained on-policy using data extracted from its own bottom-up searches on training tasks. We evaluate CrossBeam in two very different domains, string manipulation and logic programming. We observe that CrossBeam learns to search efficiently, exploring much smaller portions of the program space compared to the state-of-the-art.

PEARL: Data Synthesis via Private Embeddings and Adversarial Reconstruction Learning

- Seng Pei Liew, Tsubasa Takahashi, Michihiko Ueno
- abstract@open-review(Poster): We propose a new framework of synthesizing data using deep generative models in a differentially private manner. Within our framework, sensitive data are sanitized with rigorous privacy guarantees in a one-shot fashion, such that training deep generative models is possible without re-using the original data. Hence, no extra privacy costs or model constraints are incurred, in contrast to popular gradient sanitization approaches, which, among other issues, cause degradation in privacy guarantees as the training iteration increases. We demonstrate a realization of our framework by making use of the characteristic function and an adversarial re-weighting objective, which are of independent interest as well. Our proposal has theoretical guarantees of performance, and empirical evaluations on multiple datasets show that our approach outperforms other methods at reasonable levels of privacy.

Divisive Feature Normalization Improves Image Recognition Performance in AlexNet

- Michelle Miller, SueYeon Chung, Kenneth D. Miller
- abstract@open-review(Poster): Local divisive normalization provides a phenomenological description of many nonlinear response properties of neurons across visual cortical areas. To gain insight into the utility of this operation, we studied the effects on AlexNet of a local divisive normalization between features, with learned parameters. Developing features were arranged in a line topology, with the influence between features determined by an exponential function of the distance between them. We compared an AlexNet model with no normalization or with canonical normalizations (Batch, Group, Layer) to the same models with divisive normalization added. Divisive normalization always improved performance for models with batch or group or no normalization, generally by 1-2 percentage points, on both the CIFAR-100 and ImageNet databases. To gain insight into mechanisms underlying the improved performance, we examined several aspects of network representations. In the early layers both canonical and divisive normalizations reduced manifold capacities and increased average dimension of the individual categorical manifolds. In later layers the capacity was higher and manifold dimension lower for models roughly in order of their performance improvement. Examining the sparsity of activations across a given layer, divisive normalization layers increased sparsity, while the canonical normalization layers decreased it. Nonetheless, in the final layer, the sparseness of activity increased in the order of no normalization, divisive, com- bined, and canonical. We also investigated how the receptive fields (RFs) in the first convolutional layer (where RFs are most interpretable) change with normalization. Divisive normalization enhanced RF Fourier power at low wavelengths, while divisive+canonical enhanced power at mid (batch, group) or low (layer) wavelengths, compared to canonical alone or no normalization. In conclusion, divisive normalization enhances image recognition performance, most strongly when combined with canonical

normalization, and in doing so it reduces manifold capacity and sparsity in early layers while increasing them in final layers, and increases low- or mid-wavelength power in the first-layer receptive fields.

Evaluating Distributional Distortion in Neural Language Modeling

- Benjamin LeBrun, Alessandro Sordoni, Timothy J. O'Donnell
- abstract@open-review(Poster): A fundamental characteristic of natural language is the high rate at which speakers produce novel expressions. Because of this novelty, a heavy-tail of rare events accounts for a significant amount of the total probability mass of distributions in language (Baayen, 2001). Standard language modeling metrics such as perplexity quantify the performance of language models (LM) in aggregate. As a result, we have relatively little understanding of whether neural LMs accurately estimate the probability of sequences in this heavy-tail of rare events. To address this gap, we develop a controlled evaluation scheme which uses generative models trained on natural data as artificial languages from which we can exactly compute sequence probabilities. Training LMs on generations from these artificial languages, we compare the sequence-level probability estimates given by LMs to the true probabilities in the target language. Our experiments reveal that LSTM and Transformer language models (i) systematically underestimate the probability of sequences drawn from the target language, and (ii) do so more severely for less-probable sequences. Investigating where this probability mass went, (iii) we find that LMs tend to overestimate the probability of ill formed (perturbed) sequences. In addition, we find that this underestimation behaviour (iv) is weakened, but not eliminated by greater amounts of training data, and (v) is exacerbated for target distributions with lower entropy.

MaGNET: Uniform Sampling from Deep Generative Network Manifolds Without Retraining

- Ahmed Imtiaz Humayun, Randall Balestriero, Richard Baraniuk
- abstract@open-review(Poster): Deep Generative Networks (DGNs) are extensively employed in Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and their variants to approximate the data manifold, and data distribution on that manifold. However, training samples are often obtained based on preferences, costs, or convenience producing artifacts in the empirical data distribution e.g. the large fraction of smiling faces in the CelebA dataset or the large fraction of dark-haired individuals in FFHQ). {\em These inconsistencies will be reproduced when sampling from the trained DGN, which has far-reaching potential implications for fairness, data augmentation, anomaly detection, domain adaptation, and beyond.} In response, we develop a differential geometry based sampler -coined MaGNET- that, given any trained DGN, produces samples that are uniformly distributed on the learned manifold. We prove theoretically and empirically that our technique produces a uniform distribution on the manifold regardless of the training set distribution. We perform a range of experiments on various datasets and DGNs. One of them considers the state-of-the-art StyleGAN2 trained on FFHQ dataset, where uniform sampling via MaGNET increases distribution precision \& recall by 4.12\% \& 3.01\% and decreases gender bias by 41.2\%, without requiring labels or retraining.

Neural Contextual Bandits with Deep Representation and Shallow Exploration

- Pan Xu, Zheng Wen, Handong Zhao, Quanquan Gu
- abstract@open-review(Poster): We study neural contextual bandits, a general class of contextual bandits, where each context-action pair is associated with a raw feature vector, but the specific reward generating function is unknown. We propose a novel learning algorithm that transforms the raw feature vector using the last hidden layer of a deep ReLU neural network (deep representation learning), and uses an upper confidence bound (UCB) approach to explore in the last linear layer (shallow exploration). We prove that under standard assumptions, our proposed algorithm achieves $\tilde{O}(\sqrt{T})$ finite-time regret, where T is the learning time horizon. Compared with existing neural contextual bandit algorithms, our approach is computationally much more efficient since it only needs to explore in the last layer of the deep neural network.

PI3NN: Out-of-distribution-aware Prediction Intervals from Three Neural Networks

- Siyan Liu, Pei Zhang, Dan Lu, Guannan Zhang
- abstract@open-review(Poster): We propose a novel prediction interval (PI) method for uncertainty quantification, which addresses three major issues with the state-of-the-art PI methods. First, existing PI methods require retraining of neural networks (NNs) for every given confidence level and suffer from the crossing issue in calculating multiple PIs. Second, they usually rely on customized loss functions with extra sensitive hyperparameters for which fine tuning is required to achieve a well-calibrated PI. Third, they usually underestimate uncertainties of out-of-distribution (OOD) samples leading to over-confident PIs. Our PI3NN method calculates PIs from linear combinations of three NNs, each of which is independently trained using the standard mean squared error loss. The coefficients of the linear combinations are computed using root-finding algorithms to ensure tight PIs for a given confidence level. We theoretically prove that PI3NN can calculate PIs for a series of confidence levels without retraining NNs and it completely avoids the crossing issue. Additionally, PI3NN does not introduce any unusual hyperparameters resulting in a stable performance. Furthermore, we address OOD identification challenge by introducing an initialization scheme which provides reasonably larger PIs of the OOD samples than those of the in-distribution samples. Benchmark and real-world experiments show that our method outperforms several state-of-the-art approaches with respect to predictive uncertainty quality, robustness, and OOD samples identification.

Discriminative Similarity for Data Clustering

- Yingzhen Yang, Ping Li
- abstract@open-review(Poster): Similarity-based clustering methods separate data into clusters according to the pairwise similarity between the data, and the pairwise similarity is crucial for their performance. In this paper, we propose {\em Clustering by Discriminative Similarity (CDS)}, a novel method which learns discriminative similarity for data clustering. CDS learns an unsupervised similarity-based classifier from each data partition, and searches for the optimal partition of the data by minimizing the generalization error of the learnt classifiers associated with the data partitions. By generalization analysis via Rademacher complexity, the generalization error bound for the unsupervised similarity-based classifier is expressed as the sum of discriminative similarity between the data from different classes. It is proved that the derived discriminative similarity can also be induced by the integrated squared error bound for kernel density classification. In order to evaluate the performance of the proposed discriminative similarity, we propose a new clustering method using a kernel as the similarity function, CDS via unsupervised kernel classification (CDSK), with its effectiveness demonstrated by experimental results.

It Takes Four to Tango: Multiagent Self Play for Automatic Curriculum Generation

- Yuqing Du, Pieter Abbeel, Aditya Grover
- abstract@open-review(Poster): We are interested in training general-purpose reinforcement learning agents that can solve a wide variety of goals. Training such agents efficiently requires automatic generation of a goal curriculum. This is challenging as it requires (a) exploring goals of increasing difficulty, while ensuring that the agent (b) is exposed to a diverse set of goals in a sample efficient manner and (c) does not catastrophically forget previously solved goals. We propose Curriculum Self Play (CuSP), an automated goal generation framework that seeks to satisfy these desiderata by virtue of a multi-player game with 4 agents. We extend the asymmetric curricula learning in PAIRED (Dennis et al., 2020) to a symmetrized game that carefully balances cooperation and competition between two off-policy student learners and two regret-maximizing teachers. CuSP additionally introduces entropic goal coverage and accounts for the non-stationary nature of the students, allowing us to automatically induce a curriculum that balances progressive exploration with anti-catastrophic exploitation. We demonstrate that our method succeeds at generating an effective curricula of goals for a range of control tasks, outperforming other methods at zero-shot test-time generalization to novel out-of-distribution goals.

CROP: Certifying Robust Policies for Reinforcement Learning through Functional Smoothing

- Fan Wu, Linyi Li, Zijian Huang, Yevgeniy Vorobeychik, Ding Zhao, Bo Li
- abstract@open-review(Poster): As reinforcement learning (RL) has achieved great success and been even adopted in safety-critical domains such as autonomous vehicles, a range of empirical studies have been conducted to improve its robustness against adversarial attacks. However, how to certify its robustness with theoretical guarantees still remains challenging. In this paper, we present the ï¬ rst uniï¬ ed framework CROP (Certifying Robust Policies for RL) to provide robustness certiï¬ cation on both action and reward levels. In particular, we propose two robustness certiï¬ cation criteria: robustness of per-state actions and lower bound of cumulative rewards. We then develop a local smoothing algorithm for policies derived from Q-functions to guarantee the robustness of actions taken along the trajectory; we also develop a global smoothing algorithm for certifying the lower bound of a ï¬ nite-horizon cumulative reward, as well as a novel local smoothing algorithm to perform adaptive search in order to obtain tighter reward certiï¬ cation. Empirically, we apply CROP to evaluate several existing empirically robust RL algorithms, including adversarial training and different robust regularization, in four environments (two representative Atari games, Highway, and CartPole). Furthermore, by evaluating these algorithms against adversarial attacks, we demonstrate that our certiï¬ cations are often tight. All experiment results are available at website https://crop-leaderboard.github.io.

Neural Link Prediction with Walk Pooling

- Liming Pan, Cheng Shi, Ivan DokmaniÄ‡
- abstract@open-review(Poster): Graph neural networks achieve high accuracy in link prediction by jointly leveraging graph topology and node attributes. Topology, however, is represented indirectly; state-of-the-art methods based on subgraph classification label nodes with distance to the target link, so that, although topological information is present, it is tempered by pooling. This makes it challenging to leverage features like loops and motifs associated with network formation mechanisms. We propose a link prediction algorithm based on a new pooling scheme called WalkPool. WalkPool combines the expressivity of topological heuristics with the feature-learning ability of neural networks. It summarizes a putative link by random walk probabilities of adjacent paths. Instead of extracting transition probabilities from the original graph, it computes the transition matrix of a ``predictive" latent graph by applying attention to learned features; this may be interpreted as feature-sensitive topology fingerprinting. WalkPool can leverage unsupervised node features or be combined with GNNs and trained end-to-end. It outperforms state-of-the-art methods on all common link prediction benchmarks, both homophilic and heterophilic, with and without node attributes. Applying WalkPool to a set of unsupervised GNNs significantly improves prediction accuracy, suggesting that it may be used as a general-purpose graph pooling scheme.

On the Convergence of Certified Robust Training with Interval Bound Propagation

- Yihan Wang, Zhouxing Shi, Quanquan Gu, Cho-Jui Hsieh
- abstract@open-review(Poster): Interval Bound Propagation (IBP) is so far the base of state-of-the-art methods for training neural networks with certifiable robustness guarantees when potential adversarial perturbations present, while the convergence of IBP training remains unknown in existing literature. In this paper, we present a theoretical analysis on the convergence of IBP training. With an overparameterized assumption, we analyze the convergence of IBP robust training. We show that when using IBP training to train a randomly initialized two-layer ReLU neural network with logistic loss, gradient descent can linearly converge to zero robust training error with a high probability if we have sufficiently small perturbation radius and large network width.

Pretraining Text Encoders with Adversarial Mixture of Training Signal Generators

- Yu Meng, Chenyan Xiong, Payal Bajaj, saurabh tiwary, Paul N. Bennett, Jiawei Han, Xia Song
- abstract@open-review(Poster): We present a new framework AMOS that pretrains text encoders with an Adversarial learning curriculum via a Mixture Of Signals from multiple auxiliary generators. Following ELECTRA-style pretraining, the main encoder is trained as a discriminator to detect replaced tokens generated by auxiliary masked language models (MLMs). Different from ELECTRA which trains one MLM as the generator, we jointly train multiple MLMs of different sizes to provide training signals at various levels of difficulty. To push the discriminator to learn better with challenging replaced tokens, we learn mixture weights over the auxiliary MLMs' outputs to maximize the discriminator loss by backpropagating the gradient from the discriminator via Gumbel-Softmax. For better pretraining efficiency, we propose a way to assemble multiple MLMs into one unified auxiliary model. AMOS outperforms ELECTRA and recent state-of-the-art pretrained models by about 1 point on the GLUE benchmark for BERT base-sized models.

Towards Training Billion Parameter Graph Neural Networks for Atomic Simulations

- Anuroop Sriram, Abhishek Das, Brandon M Wood, Siddharth Goyal, C. Lawrence Zitnick
- abstract@open-review(Poster): Recent progress in Graph Neural Networks (GNNs) for modeling atomic simulations has the potential to revolutionize catalyst discovery, which is a key step in making progress towards the energy breakthroughs needed to combat climate change. However, the GNNs that have proven most effective for this task are memory intensive as they model higher-order interactions in the graphs such as those between triplets or quadruplets of atoms, making it challenging to scale these models. In this paper, we introduce Graph Parallelism, a method to distribute input graphs across multiple GPUs, enabling us to train very large GNNs with hundreds of millions or billions of parameters. We empirically evaluate our method by scaling up the recently proposed DimeNet++ and GemNet models by over an order of magnitude in the number of parameters. On the large-scale Open Catalyst 2020 (OC20) dataset, these graph-parallelized models lead to relative improvements of 1) 15% on the force MAE metric on the S2EF task and 2) 21% on the AFbT metric on the IS2RS task, establishing new state-of-the-art results.

Understanding and Leveraging Overparameterization in Recursive Value Estimation

- Chenjun Xiao, Bo Dai, Jincheng Mei, Oscar A Ramirez, Ramki Gummadi, Chris Harris, Dale Schuurmans
- abstract@open-review(Poster): The theory of function approximation in reinforcement learning (RL) typically considers low capacity representations that incur a tradeoff between approximation error, stability and generalization. Current deep architectures, however, operate in an overparameterized regime where approximation error is not necessarily a bottleneck. To better understand the utility of deep models in RL we present an analysis of recursive value estimation using \emph{overparameterized} linear representations that provides useful, transferable findings. First, we show that classical updates such as temporal difference (TD) learning or fitted-value-iteration (FVI) converge to \emph{different} fixed points than residual minimization (RM) in the overparameterized linear case. We then develop a unified interpretation of overparameterized linear value estimation as minimizing the Euclidean norm of the weights subject to alternative constraints. A practical consequence is that RM can be modified by a simple alteration of the backup targets to obtain the same fixed points as FVI and TD (when they converge), while universally ensuring stability. Further, we provide an analysis of the generalization error of these methods, demonstrating per iterate bounds on the value prediction error of FVI, and fixed point bounds for TD and RM.
  Given this understanding, we then develop new algorithmic tools for improving recursive value estimation with deep models. In particular, we extract two regularizers that penalize out-of-span top-layer weights and co-linearity in top-layer features respectively. Empirically we find that these regularizers dramatically improve the stability of TD and FVI, while allowing RM to match and even sometimes surpass their generalization performance with assured stability.

Optimization and Adaptive Generalization of Three layer Neural Networks

- Khashayar Gatmiry, Stefanie Jegelka, Jonathan Kelner
- abstract@open-review(Poster): While there has been substantial recent work studying generalization of neural networks, the ability of deep nets in automating the process of feature extraction still evades a thorough mathematical understanding.
  As a step toward this goal, we analyze learning and generalization of a three-layer neural network with ReLU activations in a regime that goes beyond the linear approximation of the network, and is hence not captured by the common Neural Tangent Kernel. We show that despite nonconvexity of the empirical loss, a variant of SGD converges in polynomially many iterations to a good solution that generalizes. In particular, our generalization bounds are adaptive: they automatically optimize over a family of kernels that includes the Neural Tangent Kernel, to provide the tightest bound.

Non-Parallel Text Style Transfer with Self-Parallel Supervision

- Ruibo Liu, Chongyang Gao, Chenyan Jia, Guangxuan Xu, Soroush Vosoughi
- abstract@open-review(Poster): The performance of existing text style transfer models is severely limited by the non-parallel datasets on which the models are trained. In non-parallel datasets, no direct mapping exists between sentences of the source and target style; the style transfer models thus only receive weak supervision of the target sentences during training, which often leads the model to discard too much style-independent information, or utterly fail to transfer the style.

In this work, we propose LaMer, a novel text style transfer framework based on large-scale language models. LaMer first mines the roughly parallel expressions in the non-parallel datasets with scene graphs, and then employs MLE training, followed by imitation learning refinement, to leverage the intrinsic parallelism within the data. On two benchmark tasks (sentiment & formality transfer) and a newly proposed challenging task (political stance transfer), our model achieves qualitative advances in transfer accuracy, content preservation, and fluency. Further empirical and human evaluations demonstrate that our model not only makes training more efficient, but also generates more readable and diverse expressions than previous models.

Can an Image Classifier Suffice For Action Recognition?

- Quanfu Fan, Chun-Fu Chen, Rameswar Panda
- abstract@open-review(Poster): We explore a new perspective on video understanding by casting the video recognition problem as an image recognition task. Our approach rearranges input video frames into super images, which allow for training an image classifier directly to fulfill the task of action recognition, in exactly the same way as image classification. With such a simple idea, we show that transformer-based image classifiers alone can suffice for action recognition. In particular, our approach demonstrates strong and promising performance against SOTA methods on several public datasets including Kinetics400, Moments In Time, Something-Something V2 (SSV2), Jester and Diving48. We also experiment with the prevalent ResNet image classifiers in computer vision to further validate our idea. The results on both Kinetics400 and SSV2 are comparable to some of the best-performed CNN approaches based on spatio-temporal modeling. Our source codes and models are available at \url{https://github.com/IBM/sifar-pytorch}.

Interacting Contour Stochastic Gradient Langevin Dynamics

- Wei Deng, Siqi Liang, Botao Hao, Guang Lin, Faming Liang
- abstract@open-review(Poster): We propose an interacting contour stochastic gradient Langevin dynamics (ICSGLD) sampler, an embarrassingly parallel multiple-chain contour stochastic gradient Langevin dynamics (CSGLD) sampler with efficient interactions. We show that ICSGLD can be theoretically more efficient than a single-chain CSGLD with an equivalent computational budget. We also present a novel random-field function, which facilitates the estimation of self-adapting parameters in big data and obtains free mode explorations. Empirically, we compare the proposed algorithm with popular benchmark methods for posterior sampling. The numerical results show a great potential of ICSGLD for large-scale uncertainty estimation tasks.

NeuPL: Neural Population Learning

- Siqi Liu, Luke Marris, Daniel Hennes, Josh Merel, Nicolas Heess, Thore Graepel
- abstract@open-review(Poster): Learning in strategy games (e.g. StarCraft, poker) requires the discovery of diverse policies. This is often achieved by iteratively training new policies against existing ones, growing a policy population that is robust to exploit. This iterative approach suffers from two issues in real-world games: a) under finite budget, approximate best-response operators at each iteration needs truncating, resulting in under-trained good-responses populating the population; b) repeated learning of basic skills at each iteration is wasteful and becomes intractable in the presence of increasingly strong opponents. In this work, we propose Neural Population Learning (NeuPL) as a solution to both issues. NeuPL offers convergence guarantees to a population of best-responses under mild assumptions. By representing a population of policies within a single conditional model, NeuPL enables transfer learning across policies. Empirically, we show the generality, improved performance and efficiency of NeuPL across several test domains. Most interestingly, we show that novel strategies become more accessible, not less, as the neural population expands.

DeSKO: Stability-Assured Robust Control with a Deep Stochastic Koopman Operator

- Minghao Han, Jacob Euler-Rolle, Robert K. Katzschmann
- abstract@open-review(Poster): The Koopman operator theory linearly describes nonlinear dynamical systems in a high-dimensional functional space and it allows to apply linear control methods to highly nonlinear systems. However, the Koopman operator does not account for any uncertainty in dynamical systems, causing it to perform poorly in real-world applications. Therefore, we propose a deep stochastic Koopman operator (DeSKO) model in a robust learning control framework to guarantee stability of nonlinear stochastic systems. The DeSKO model captures a dynamical system's uncertainty by inferring a distribution of observables. We use the inferred distribution to design a robust, stabilizing closed-loop controller for a dynamical system. Modeling and control experiments on several advanced control benchmarks show that our framework is more robust and scalable than state-of-the-art deep Koopman operators and reinforcement learning methods. Tested control benchmarks include a soft robotic arm, a legged robot, and a biological gene regulatory network. We also demonstrate that this robust control method resists previously unseen uncertainties, such as external disturbances, with a magnitude of up to five times the maximum control input. Our approach opens up new possibilities in learning control for high-dimensional nonlinear systems while robustly managing internal or external uncertainty.

Neural Network Approximation based on Hausdorff distance of Tropical Zonotopes

- Panagiotis Misiakos, Georgios Smyrnis, George Retsinas, Petros Maragos
- abstract@open-review(Poster): In this work we theoretically contribute to neural network approximation by providing a novel tropical geometrical viewpoint to structured neural network compression. In particular, we show that the approximation error between two neural networks with ReLU activations and one hidden layer depends on the Hausdorff distance of the tropical zonotopes of the networks. This theorem comes as a first step towards a purely geometrical interpretation of neural network approximation. Based on this theoretical contribution, we propose geometrical methods that employ the K-means algorithm to compress the fully connected parts of ReLU activated deep neural networks. We analyze the error bounds of our algorithms theoretically based on our approximation theorem and evaluate them empirically on neural network compression. Our experiments follow a proof-of-concept strategy and indicate that our geometrical tools achieve improved performance over relevant tropical geometry techniques and can be competitive against non-tropical methods.

Learning Towards The Largest Margins

- Xiong Zhou, Xianming Liu, Deming Zhai, Junjun Jiang, Xin Gao, Xiangyang Ji
- abstract@open-review(Poster): One of the main challenges for feature representation in deep learning-based classification is the design of appropriate loss functions that exhibit strong discriminative power. The classical softmax loss does not explicitly encourage discriminative learning of features. A popular direction of research is to incorporate margins in well-established losses in order to enforce extra intra-class compactness and inter-class separability, which, however, were developed through heuristic means, as opposed to rigorous mathematical principles. In this work, we attempt to address this limitation by formulating the principled optimization objective as learning towards the largest margins. Specifically, we firstly propose to employ the class margin as the measure of inter-class separability, and the sample margin as the measure of intra-class compactness. Accordingly, to encourage discriminative representation of features, the loss function should promote the largest possible margins for both classes and samples. Furthermore, we derive a generalized margin softmax loss to draw general conclusions for the existing margin-based losses. Not only does this principled framework offer new perspectives to understand and interpret existing margin-based losses, but it also provides new insights that can guide the design of new tools, including \textit{sample margin regularization} and \textit{largest margin softmax loss} for class balanced cases, and \textit{zero centroid regularization} for class imbalanced cases. Experimental results demonstrate the effectiveness of our strategy for multiple tasks including visual classification, imbalanced classification, person re-identification, and face verification.

Patch-Fool: Are Vision Transformers Always Robust Against Adversarial Perturbations?

- Yonggan Fu, Shunyao Zhang, Shang Wu, Cheng Wan, Yingyan Lin
- abstract@open-review(Poster): Vision transformers (ViTs) have recently set off a new wave in neural architecture design thanks to their record-breaking performance in various vision tasks. In parallel, to fulfill the goal of deploying ViTs into real-world vision applications, their robustness against potential malicious attacks has gained increasing attention. In particular, recent works show that ViTs are more robust against adversarial attacks as compared with convolutional neural networks (CNNs), and conjecture that this is because ViTs focus more on capturing global interactions among different input/feature patches, leading to their improved robustness to local perturbations imposed by adversarial attacks. In this work, we ask an intriguing question: "Under what kinds of perturbations do ViTs become more vulnerable learners compared to CNNs?" Driven by this question, we first conduct a comprehensive experiment regarding the robustness of both ViTs and CNNs under various existing adversarial attacks to understand the underlying reason favoring their robustness. Based on the drawn insights, we then propose a dedicated attack framework, dubbed Patch-Fool, that fools the self-attention mechanism by attacking its basic component (i.e., a single patch) with a series of attention-aware optimization techniques. Interestingly, our Patch-Fool framework shows for the first time that ViTs are not necessarily more robust than CNNs against adversarial perturbations. In particular, we find that ViTs are more vulnerable learners compared with CNNs against our Patch-Fool attack which is consistent across extensive experiments, and the observations from Sparse/Mild Patch-Fool, two variants of Patch-Fool, indicate an intriguing insight that the perturbation density and strength on each patch seem to be the key factors that influence the robustness ranking between ViTs and CNNs. It can be expected that our Patch-Fool framework will shed light on both future architecture designs and training schemes for robustifying ViTs towards their real-world deployment. Our codes are available at https://github.com/RICE-EIC/Patch-Fool.

AdaMatch: A Unified Approach to Semi-Supervised Learning and Domain Adaptation

- David Berthelot, Rebecca Roelofs, Kihyuk Sohn, Nicholas Carlini, Alexey Kurakin
- abstract@open-review(Poster): We extend semi-supervised learning to the problem of domain adaptation to learn significantly higher-accuracy models that train on one data distribution and test on a different one. With the goal of generality, we introduce AdaMatch, a unified solution for unsupervised domain adaptation (UDA), semi-supervised learning (SSL), and semi-supervised domain adaptation (SSDA). In an extensive experimental study, we compare its behavior with respective state-of-the-art techniques from SSL, SSDA, and UDA and find that AdaMatch either matches or significantly exceeds the state-of-the-art in each case using the same hyper-parameters regardless of the dataset or task. For example, AdaMatch nearly doubles the accuracy compared to that of the prior state-of-the-art on the UDA task for DomainNet and even exceeds the accuracy of the prior state-of-the-art obtained with pre-training by 6.4% when AdaMatch is trained completely from scratch. Furthermore, by providing AdaMatch with just one labeled example per class from the target domain (i.e., the SSDA setting), we increase the target accuracy by an additional 6.1%, and with 5 labeled examples, by 13.6%.

Complete Verification via Multi-Neuron Relaxation Guided Branch-and-Bound

- Claudio Ferrari, Mark Niklas Mueller, Nikola JovanoviÄ‡, Martin Vechev
- abstract@open-review(Poster): State-of-the-art neural network verifiers are fundamentally based on one of two paradigms: either encoding the whole verification problem via tight multi-neuron convex relaxations or applying a Branch-and-Bound (BaB) procedure leveraging imprecise but fast bounding methods on a large number of easier subproblems. The former can capture complex multi-neuron dependencies but sacrifices completeness due to the inherent limitations of convex relaxations. The latter enables complete verification but becomes increasingly ineffective on larger and more challenging networks. In this work, we present a novel complete verifier which combines the strengths of both paradigms: it leverages multi-neuron relaxations to drastically reduce the number of subproblems generated during the BaB process and an efficient GPU-based dual optimizer to solve the remaining ones. An extensive evaluation demonstrates that our verifier achieves a new state-of-the-art on both established benchmarks as well as networks with significantly higher accuracy than previously considered. The latter result (up to 28% certification gains) indicates meaningful progress towards creating verifiers that can handle practically relevant networks.

Learning Fast Samplers for Diffusion Models by Differentiating Through Sample Quality

- Daniel Watson, William Chan, Jonathan Ho, Mohammad Norouzi
- abstract@open-review(Poster): Diffusion models have emerged as an expressive family of generative models rivaling GANs in sample quality and autoregressive models in likelihood scores. Standard diffusion models typically require hundreds of forward passes through the model to generate a single high-fidelity sample. We introduce Differentiable Diffusion Sampler Search (DDSS): a method that optimizes fast samplers for any pre-trained diffusion model by differentiating through sample quality scores. We also present Generalized Gaussian Diffusion Models (GGDM), a family of flexible non-Markovian samplers for diffusion models. We show that optimizing the degrees of freedom of GGDM samplers by maximizing sample quality scores via gradient descent leads to improved sample quality. Our optimization procedure backpropagates through the sampling process using the reparametrization trick and gradient rematerialization. DDSS achieves strong results on unconditional image generation across various datasets (e.g., FID scores on LSUN church 128x128 of 11.6 with only 10 inference steps, and 4.82 with 20 steps, compared to 51.1 and 14.9 with strongest DDPM/DDIM baselines). Our method is compatible with any pre-trained diffusion model without fine-tuning or re-training required.

Distribution Compression in Near-Linear Time

- Abhishek Shetty, Raaz Dwivedi, Lester Mackey
- abstract@open-review(Poster): In distribution compression, one aims to accurately summarize a probability distribution $\mathbb{P}$ using a small number of representative points. Near-optimal thinning procedures achieve this goal by sampling n points from a Markov chain and identifying $\sqrt{n}$ points with $\widetilde{\mathcal{O}}(1/\sqrt{n})$ discrepancy to $\mathbb{P}$. Unfortunately, these algorithms suffer from quadratic or super-quadratic runtime in the sample size n. To address this deficiency, we introduce Compress++, a simple meta-procedure for speeding up any thinning algorithm while suffering at most a factor of 4 in error. When combined with the quadratic-time kernel halving and kernel thinning algorithms of Dwivedi and Mackey (2021), Compress++ delivers $\sqrt{n}$ points with $\mathcal{O}(\sqrt{\log n/n})$ integration error and better-than-Monte-Carlo maximum mean discrepancy in $\mathcal{O}(n \log^3 n)$ time and $\mathcal{O}(\sqrt{n} \log^2 n)$ space. Moreover, Compress++ enjoys the same near-linear runtime given any quadratic-time input and reduces the runtime of super-quadratic algorithms by a square-root factor. In our benchmarks

with high-dimensional Monte Carlo samples and Markov chains targeting challenging differential equation posteriors, Compress++ matches or nearly matches the accuracy of its input algorithm in orders of magnitude less time.

Capturing Structural Locality in Non-parametric Language Models

- Frank F. Xu, Junxian He, Graham Neubig, Vincent Josua Hellendoorn
- abstract@open-review(Poster): Structural locality is a ubiquitous feature of real-world datasets, wherein data points are organized into local hierarchies. Some examples include topical clusters in text or project hierarchies in source code repositories. In this paper, we explore utilizing this structural locality within non-parametric language models, which generate sequences that reference retrieved examples from an external source. We propose a simple yet effective approach for adding locality information into such models by adding learned parameters that improve the likelihood of retrieving examples from local neighborhoods. Experiments on two different domains, Java source code and Wikipedia text, demonstrate that locality features improve model efficacy over models without access to these features, with interesting differences. We also perform an analysis of how and where locality features contribute to improving performance and why the traditionally used contextual similarity metrics alone are not enough to grasp the locality structure.

Audio Lottery: Speech Recognition Made Ultra-Lightweight, Noise-Robust, and Transferable

- Shaojin Ding, Tianlong Chen, Zhangyang Wang
- abstract@open-review(Poster): Lightweight speech recognition models have seen explosive demands owing to a growing amount of speech-interactive features on mobile devices. Since designing such systems from scratch is non-trivial, practitioners typically choose to compress large (pre-trained) speech models. Recently, lottery ticket hypothesis reveals the existence of highly sparse subnetworks that can be trained in isolation without sacrificing the performance of the full models. In this paper, we investigate the tantalizing possibility of using lottery ticket hypothesis to discover lightweight speech recognition models, that are (1) robust to various noise existing in speech; (2) transferable to fit the open-world personalization; and 3) compatible with structured sparsity. We conducted extensive experiments on CNN-LSTM, RNN-Transducer, and Transformer models, and verified the existence of highly sparse winning tickets that can match the full model performance across those backbones. We obtained winning tickets that have less than 20% of full model weights on all backbones, while the most lightweight one only keeps 4.4% weights. Those winning tickets generalize to structured sparsity with no performance loss, and transfer exceptionally from large source datasets to various target datasets. Perhaps most surprisingly, when the training utterances have high background noises, the winning tickets even substantially outperform the full models, showing the extra bonus of noise robustness by inducing sparsity. Codes are available at https://github.com/VITA-Group/Audio-Lottery.

Learning to Map for Active Semantic Goal Navigation

- Georgios Georgakis, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Kostas Daniilidis
- abstract@open-review(Poster): We consider the problem of object goal navigation in unseen environments. Solving this problem requires learning of contextual semantic priors, a challenging endeavour given the spatial and semantic variability of indoor environments. Current methods learn to implicitly encode these priors through goal-oriented navigation policy functions operating on spatial representations that are limited to the agent's observable areas. In this work, we propose a novel framework that actively learns to generate semantic maps outside the field of view of the agent and leverages the uncertainty over the semantic classes in the unobserved areas to decide on long term goals. We demonstrate that through this spatial prediction strategy, we are able to learn semantic priors in scenes that can be leveraged in unknown environments. Additionally, we show how different objectives can be defined by balancing exploration with exploitation during searching for semantic targets. Our method is validated in the visually realistic environments of the Matterport3D dataset and show improved results on object goal navigation over competitive baselines.

Benchmarking the Spectrum of Agent Capabilities

- Danijar Hafner
- abstract@open-review(Poster): Evaluating the general abilities of intelligent agents requires complex simulation environments. Existing benchmarks typically evaluate only one narrow task per environment, requiring researchers to perform expensive training runs on many different environments. We introduce Crafter, an open world survival game with visual inputs that evaluates a wide range of general abilities within a single environment. Agents either learn from the provided reward signal or through intrinsic objectives and are evaluated by semantically meaningful achievements that can be unlocked during each episode, such as discovering resources and crafting tools. Consistently unlocking all achievements requires strong generalization, deep exploration, and long-term reasoning. We experimentally verify that Crafter is of appropriate difficulty to drive future research and provide baselines scores of reward agents and unsupervised agents. Furthermore, we observe sophisticated behaviors emerging from maximizing the reward signal, such as building tunnel systems, bridges, houses, and plantations. We hope that Crafter will accelerate research progress by quickly evaluating a wide spectrum of abilities.

Mind the Gap: Domain Gap Control for Single Shot Domain Adaptation for Generative Adversarial Networks

- Peihao Zhu, Rameen Abdal, John Femiani, Peter Wonka
- abstract@open-review(Poster): We present a new method for one shot domain adaptation. The input to our method is trained GAN that can produce images in domain A and a single reference image I_B from domain B. The proposed algorithm can translate any output of the trained GAN from domain A to domain B. There are two main advantages of our method compared to the current state of the art: First, our solution achieves higher visual quality, e.g. by noticeably reducing overfitting. Second, our solution allows for more degrees of freedom to control the domain gap, i.e. what aspects of image I_B are used to define the domain B. Technically, we realize the new method by building on a pre-trained StyleGAN generator as GAN and a pre-trained CLIP model for representing the domain gap. We propose several new regularizers for controlling the domain gap to optimize the weights of the pre-trained StyleGAN generator to output images in domain B instead of domain A. The regularizers prevent the optimization from taking on too many attributes of the single reference image. Our results show significant visual improvements over the state of the art as well as multiple applications that highlight improved control.

On Evaluation Metrics for Graph Generative Models

- Rylee Thompson, Boris Knyazev, Elahe Ghalebi, Jungtaek Kim, Graham W. Taylor
- abstract@open-review(Poster): In image generation, generative models can be evaluated naturally by visually inspecting model outputs. However, this is not always the case for graph generative models (GGMs), making their evaluation challenging. Currently, the standard process for evaluating GGMs suffers from three critical limitations: i) it does not produce a single score which makes model selection challenging, ii) in many cases it fails to consider underlying edge and node features, and iii) it is prohibitively slow to perform. In this work, we mitigate these issues by searching for \emph{scalar, domain-agnostic, and scalable metrics} for evaluating and ranking GGMs. To this end, we study existing GGM metrics and neural-network-based metrics emerging from generative models of images that use embeddings extracted from a task-specific network. Motivated by the power of Graph Neural Networks (GNNs) to extract meaningful graph representations \emph{without any training}, we introduce several metrics based on the features extracted by an untrained random GNN. We design experiments to thoroughly test and objectively score metrics on their ability to measure the diversity and fidelity of generated graphs, as well as their sample and computational efficiency. Depending on the quantity of samples, we recommend one of two metrics from our collection of random-GNN-based metrics. We show these two metrics to be more expressive than pre-existing and alternative random-GNN-based metrics using our objective scoring. While we focus on applying these metrics to GGM evaluation, in practice this enables the ability to easily compute the dissimilarity between any two sets of graphs \emph{regardless of domain}. Our code is released at: https://github.com/uoguelph-mlrg/GGM-metrics.

Selective Ensembles for Consistent Predictions

- Emily Black, Klas Leino, Matt Fredrikson
- abstract@open-review(Poster): Recent work has shown that models trained to the same objective, and which achieve similar measures of accuracy on consistent test data, may nonetheless behave very differently on individual predictions. This inconsistency is undesirable in high-stakes contexts, such as medical diagnosis and finance. We show that this duplicitous behavior extends beyond predictions to feature attributions, which may likewise have negative implications for the intelligibility of a model, and one's ability to find recourse for subjects. We then introduce selective ensembles to mitigate such inconsistencies by applying hypothesis testing to the predictions of a set of models trained using randomly-selected starting conditions; importantly, selective ensembles can abstain in cases where a consistent outcome cannot be achieved up to a specified confidence level. We prove that that prediction disagreement between selective ensembles is bounded, and empirically demonstrate that selective ensembles achieve consistent predictions and feature attributions while maintaining low abstention rates. On several benchmark datasets, selective ensembles reach zero inconsistently predicted points, with abstention rates as low as 1.5%.

Graph Condensation for Graph Neural Networks

- Wei Jin, Lingxiao Zhao, Shichang Zhang, Yozen Liu, Jiliang Tang, Neil Shah
- abstract@open-review(Poster): Given the prevalence of large-scale graphs in real-world applications, the storage and time for training neural models have raised increasing concerns. To alleviate the concerns, we propose and study the problem of graph condensation for graph neural networks (GNNs). Specifically, we aim to condense the large, original graph into a small, synthetic and highly-informative graph, such that GNNs trained on the small graph and large graph have comparable performance. We approach the condensation problem by imitating the GNN training trajectory on the original graph through the optimization of a gradient matching loss and design a strategy to condense node futures and structural information simultaneously. Extensive experiments have demonstrated the effectiveness of the proposed framework in condensing different graph datasets into informative smaller graphs. In particular, we are able to approximate the original test accuracy by 95.3\% on Reddit, 99.8\% on Flickr and 99.0\% on Citeseer, while reducing their graph size by more than 99.9\%, and the condensed graphs can be used to train various GNN architectures.

DIVA: Dataset Derivative of a Learning Task

- Yonatan Dukler, Alessandro Achille, Giovanni Paolini, Avinash Ravichandran, Marzia Polito, Stefano Soatto
- abstract@open-review(Poster): We present a method to compute the derivative of a learning task with respect to a dataset. A learning task is a function from a training set to the validation error, which can be represented by a trained deep neural network (DNN). The ``dataset derivative'' is a linear operator, computed around the trained model, that informs how perturbations of the weight of each training sample affect the validation error, usually computed on a separate validation dataset. Our method, DIVA (Differentiable Validation) hinges on a closed-form differentiable expression of the leave-one-out cross-validation error around a pre-trained DNN. Such expression constitutes the dataset derivative. DIVA could be used for dataset auto-curation, for example removing samples with faulty annotations, augmenting a dataset with additional relevant samples, or rebalancing. More generally, DIVA can be used to optimize the dataset, along with the parameters of the model, as part of the training process without the need for a separate validation dataset, unlike bi-level optimization methods customary in AutoML. To illustrate the flexibility of DIVA, we report experiments on sample auto-curation tasks such as outlier rejection, dataset extension, and automatic aggregation of multi-modal data.

Towards General Function Approximation in Zero-Sum Markov Games

- Baihe Huang, Jason D. Lee, Zhaoran Wang, Zhuoran Yang
- abstract@open-review(Poster): This paper considers two-player zero-sum finite-horizon Markov games with simultaneous moves. The study focuses on the challenging settings where the value function or the model is parameterized by general function classes. Provably efficient algorithms for both decoupled and coordinated settings are developed. In the decoupled setting where the agent controls a single player and plays against an arbitrary opponent, we propose a new model-free algorithm. The sample complexity is governed by the Minimax Eluder dimensionâ€”a new dimension of the function class in Markov games. As a special case, this method improves the state-of-the-art algorithm by a $\sqrt{d}$ factor in the regret when the reward function and transition kernel are parameterized with d-dimensional linear features. In the coordinated setting where both players are controlled by the agent, we propose a model-based algorithm and a model-free algorithm. In the model-based algorithm, we prove that sample complexity can be bounded by a generalization of Witness rank to Markov games. The model-free algorithm enjoys a $\sqrt{K}$-regret upper bound where K is the number of episodes. Our algorithms are based on new techniques of alternate optimism

Exposing the Implicit Energy Networks behind Masked Language Models via Metropolis--Hastings

- Kartik Goyal, Chris Dyer, Taylor Berg-Kirkpatrick
- abstract@open-review(Poster): While recent work has shown that scores from models trained by the ubiquitous masked language modeling (MLM) objective effectively discriminate probable from improbable sequences, it is still an open question if these MLMs specify a principled probability distribution over the space of possible sequences. In this paper, we interpret MLMs as energy-based sequence models and propose two energy parametrizations derivable from the trained MLMs. In order to draw samples correctly from these models, we develop a tractable sampling scheme based on the Metropolis--Hastings Monte Carlo algorithm. In our approach, samples are proposed from the same masked conditionals used for training the masked language models, and they are accepted or rejected based on their energy values according to the target distribution. We validate the effectiveness of the proposed parametrizations by exploring the quality of samples drawn from these energy-based models for both open-ended unconditional generation and a conditional generation task of machine translation. We theoretically and empirically justify our sampling algorithm by showing that the masked conditionals on their own do not yield a Markov chain whose stationary distribution is that of our target distribution, and our approach generates higher quality samples than other recently proposed undirected generation approaches (Wang et al., 2019, Ghazvininejad et al., 2019).

ClimateGAN: Raising Climate Change Awareness by Generating Images of Floods

- Victor Schmidt, Alexandra Luccioni, MÃ©lisande Teng, Tianyu Zhang, Alexia Reynaud, Sunand Raghupathi, Gautier Cosne, Adrien Juraver, Vahe Vardanyan, Alex HernÃ¡ndez-GarcÃa, Yoshua Bengio
- abstract@open-review(Poster): Climate change is a major threat to humanity and the actions required to prevent its catastrophic consequences include changes in both policy-making and individual behaviour. However, taking action requires understanding its seemingly abstract and distant consequences. Projecting the potential impacts of extreme climate events such as flooding in familiar places can help make the impacts of climate change more concrete and encourage action. As part of a larger initiative to build a website (https://thisclimatedoesnotexist.com) that projects extreme climate events onto user-chosen photos, we present our solution to simulate photo-realistic floods on authentic images. To address this complex task in the absence of suitable data, we propose ClimateGAN, a model that leverages both simulated and real data through unsupervised domain adaptation and conditional image generation. In this paper, we describe the details of our framework, thoroughly evaluate the main components of our architecture and demonstrate that our model is capable of robustly generating photo-realistic flooding on street images.

A Comparison of Hamming Errors of Representative Variable Selection Methods

- Tracy Ke, Longlin Wang

- abstract@open-review(Poster): Lasso is a celebrated method for variable selection in linear models, but it faces challenges when the covariates are moderately or strongly correlated. This motivates alternative approaches such as using a non-convex penalty, adding a ridge regularization, or conducting a post-Lasso thresholding. In this paper, we compare Lasso with 5 other methods: Elastic net, SCAD, forward selection, thresholded Lasso, and forward backward selection. We measure their performances theoretically by the expected Hamming error, assuming that the regression coefficients are ${\it iid}$ drawn from a two-point mixture and that the Gram matrix is block-wise diagonal. By deriving the rates of convergence of Hamming errors and the phase diagrams, we obtain useful conclusions about the pros and cons of different methods.

A Program to Build E(N)-Equivariant Steerable CNNs

- Gabriele Cesa, Leon Lang, Maurice Weiler
- abstract@open-review(Poster): Equivariance is becoming an increasingly popular design choice to build data efficient neural networks by exploiting prior knowledge about the symmetries of the problem at hand. Euclidean steerable CNNs are one of the most common classes of equivariant networks. While the constraints these architectures need to satisfy are understood, existing approaches are tailored to specific (classes of) groups. No generally applicable method that is practical for implementation has been described so far. In this work, we generalize the Wigner-Eckart theorem proposed in Lang & Weiler (2020), which characterizes general G-steerable kernel spaces for compact groups G over their homogeneous spaces, to arbitrary G-spaces. This enables us to directly parameterize filters in terms of a band-limited basis on the whole space rather than on G's orbits, but also to easily implement steerable CNNs equivariant to a large number of groups. To demonstrate its generality, we instantiate our method on a variety of isometry groups acting on the Euclidean space $\mathbb{R}^3$. Our framework allows us to build $E(3)$ and $SE(3)$-steerable CNNs like previous works, but also CNNs with arbitrary $G\leq O(3)$-steerable kernels. For example, we build 3D CNNs equivariant to the symmetries of platonic solids or choose $G=SO(2)$ when working with 3D data having only azimuthal symmetries. We compare these models on 3D shapes and molecular datasets, observing improved performance by matching the model's symmetries to the ones of the data.

Minimax Optimization with Smooth Algorithmic Adversaries

- Tanner Fiez, Chi Jin, Praneeth Netrapalli, Lillian J Ratliff
- abstract@open-review(Poster): This paper considers minimax optimization $\min_x \max_y f(x, y)$ in the challenging setting where f can be both nonconvex in x and nonconcave in y. Though such optimization problems arise in many machine learning paradigms including training generative adversarial networks (GANs) and adversarially robust models, from a theoretical point of view, two fundamental issues remain: (i) the absence of simple and efficiently computable optimality notions, and (ii) cyclic or diverging behavior of existing algorithms. This paper proposes a new theoretical framework for nonconvex-nonconcave minimax optimization that addresses both of the above issues. The starting point of this paper is the observation that, under a computational budget, the max-player can not fully maximize $f(x,\cdot)$ since nonconcave maximization is NP-hard in general. So, we propose a new framework, and a corresponding algorithm, for the min-player to play against \emph{smooth algorithms} deployed by the adversary (i.e., the max-player) instead of against full maximization. Our algorithm is guaranteed to make monotonic progress (thus having no limit cycles or diverging behavior), and to find an appropriate ``stationary point'' in a polynomial number of iterations. Our framework covers practically relevant settings where the smooth algorithms deployed by the adversary are multi-step stochastic gradient ascent, and its accelerated version. We further present experimental results that confirm our theoretical findings and demonstrate the effectiveness of the proposed approach in practice on simple, conceptual settings.

On Distributed Adaptive Optimization with Gradient Compression

- Xiaoyun Li, Belhal Karimi, Ping Li
- abstract@open-review(Poster): We study COMP-AMS, a distributed optimization framework based on gradient averaging and adaptive AMSGrad algorithm. Gradient compression with error feedback is applied to reduce the communication cost in the gradient transmission process. Our convergence analysis of COMP-AMS shows that such compressed gradient averaging strategy yields same convergence rate as standard AMSGrad, and also exhibits the linear speedup effect w.r.t. the number of local workers. Compared with recently proposed protocols on distributed adaptive methods, COMP-AMS is simple and convenient. Numerical experiments are conducted to justify the theoretical findings, and demonstrate that the proposed method can achieve same test accuracy as the full-gradient AMSGrad with substantial communication savings. With its simplicity and efficiency, COMP-AMS can serve as a useful distributed training framework for adaptive methods.

Leveraging unlabeled data to predict out-of-distribution performance

- Saurabh Garg, Sivaraman Balakrishnan, Zachary Chase Lipton, Behnam Neyshabur, Hanie Sedghi
- abstract@open-review(Poster): Real-world machine learning deployments are characterized by mismatches between the source (training) and target (test) distributions that may cause performance drops. In this work, we investigate methods for predicting the target domain accuracy using only labeled source data and unlabeled target data. We propose Average Thresholded Confidence (ATC), a practical method that learns a \emph{threshold} on the model's confidence, predicting accuracy as the fraction of unlabeled examples for which model confidence exceeds that threshold. ATC outperforms previous methods across several model architectures, types of distribution shifts (e.g., due to synthetic corruptions, dataset reproduction, or novel subpopulations), and datasets (\textsc{Wilds}-FMoW, ImageNet, \breeds, CIFAR, and MNIST). In our experiments, ATC estimates target performance $2\text{--}4\times$ more accurately than prior methods. We also explore the theoretical foundations of the problem, proving that, in general, identifying the accuracy is just as hard as identifying the optimal predictor and thus, the efficacy of any method rests upon (perhaps unstated) assumptions on the nature of the shift. Finally, analyzing our method on some toy distributions, we provide insights concerning when it works.

VC dimension of partially quantized neural networks in the overparametrized regime

- Yutong Wang, Clayton Scott
- abstract@open-review(Poster): Vapnik-Chervonenkis (VC) theory has so far been unable to explain the small generalization error of overparametrized neural networks. Indeed, existing applications of VC theory to large networks obtain upper bounds on VC dimension that are proportional to the number of weights, and for a large class of networks, these upper bound are known to be tight. In this work, we focus on a class of partially quantized networks that we refer to as hyperplane arrangement neural networks (HANNs). Using a sample compression analysis, we show that HANNs can have VC dimension significantly smaller than the number of weights, while being highly expressive. In particular, empirical risk minimization over HANNs in the overparametrized regime achieves the minimax rate for classification with Lipschitz posterior class probability. We further demonstrate the expressivity of HANNs empirically. On a panel of 121 UCI datasets, overparametrized HANNs are able to match the performance of state-of-the-art full-precision models.

Optimal Representations for Covariate Shift

- Yangjun Ruan, Yann Dubois, Chris J. Maddison
- abstract@open-review(Poster): Machine learning systems often experience a distribution shift between training and testing. In this paper, we introduce a simple variational objective whose optima are exactly the set of all representations on which risk minimizers are guaranteed to be robust to any distribution shift that preserves the Bayes predictor, e.g., covariate shifts. Our objective has two components. First, a representation must remain discriminative for the task, i.e., some predictor must be able to simultaneously minimize the source and target risk. Second, the representation's marginal support needs to be the same across source and target. We make this practical by designing self-supervised objectives that only use unlabelled data and

augmentations to train robust representations. Our objectives give insights into the robustness of CLIP, and further improve CLIP's representations to achieve SOTA results on DomainBed.

Fortuitous Forgetting in Connectionist Networks

- Hattie Zhou, Ankit Vani, Hugo Larochelle, Aaron Courville
- abstract@open-review(Poster): Forgetting is often seen as an unwanted characteristic in both human and machine learning. However, we propose that forgetting can in fact be favorable to learning. We introduce forget-and-relearn as a powerful paradigm for shaping the learning trajectories of artificial neural networks. In this process, the forgetting step selectively removes undesirable information from the model, and the relearning step reinforces features that are consistently useful under different conditions. The forget-and-relearn framework unifies many existing iterative training algorithms in the image classification and language emergence literature, and allows us to understand the success of these algorithms in terms of the disproportionate forgetting of undesirable information. We leverage this understanding to improve upon existing algorithms by designing more targeted forgetting operations. Insights from our analysis provide a coherent view on the dynamics of iterative training in neural networks and offer a clear path towards performance improvements.

EigenGame Unloaded: When playing games is better than optimizing

- Ian Gemp, Brian McWilliams, Claire Vernade, Thore Graepel
- abstract@open-review(Poster): We build on the recently proposed EigenGame that views eigendecomposition as a competitive game. EigenGame's updates are biased if computed using minibatches of data, which hinders convergence and more sophisticated parallelism in the stochastic setting. In this work, we propose an unbiased stochastic update that is asymptotically equivalent to EigenGame, enjoys greater parallelism allowing computation on datasets of larger sample sizes, and outperforms EigenGame in experiments. We present applications to finding the principal components of massive datasets and performing spectral clustering of graphs. We analyze and discuss our proposed update in the context of EigenGame and the shift in perspective from optimization to games.

Contextualized Scene Imagination for Generative Commonsense Reasoning

- PeiFeng Wang, Jonathan Zamora, Junfeng Liu, Filip Ilievski, Muhao Chen, Xiang Ren
- abstract@open-review(Poster): Humans use natural language to compose common concepts from their environment into plausible, day-to-day scene descriptions. However, such generative commonsense reasoning (GCSR) skills are lacking in state-of-the-art text generation methods. Descriptive sentences about arbitrary concepts generated by neural text generation models (e.g., pre-trained text-to-text Transformers) are often grammatically fluent but may not correspond to human common sense, largely due to their lack of mechanisms to capture concept relations, to identify implicit concepts, and to perform generalizable reasoning about unseen concept compositions. In this paper, we propose an Imagine-and-Verbalize (I\&V) method, which learns to imagine a relational scene knowledge graph (SKG) with relations between the input concepts, and leverage the SKG as a constraint when generating a plausible scene description. We collect and harmonize a set of knowledge resources from different domains and modalities, providing a rich auxiliary supervision signal for I\&V. The experiments demonstrate the effectiveness of I\&V in improving language models on both concept-to-sentence and concept-to-story generation tasks, while enabling the model to learn well from fewer task examples and generate SKGs that make common sense to human annotators.

Scene Transformer: A unified architecture for predicting future trajectories of multiple agents

- Jiquan Ngiam, Vijay Vasudevan, Benjamin Caine, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, David J Weiss, Ben Sapp, Zhifeng Chen, Jonathon Shlens
- abstract@open-review(Poster): Predicting the motion of multiple agents is necessary for planning in dynamic environments. This task is challenging for autonomous driving since agents (e.g., vehicles and pedestrians) and their associated behaviors may be diverse and influence one another. Most prior work have focused on predicting independent futures for each agent based on all past motion, and planning against these independent predictions. However, planning against independent predictions can make it challenging to represent the future interaction possibilities between different agents, leading to sub-optimal planning. In this work, we formulate a model for predicting the behavior of all agents jointly, producing consistent futures that account for interactions between agents. Inspired by recent language modeling approaches, we use a masking strategy as the query to our model, enabling one to invoke a single model to predict agent behavior in many ways, such as potentially conditioned on the goal or full future trajectory of the autonomous vehicle or the behavior of other agents in the environment. Our model architecture employs attention to combine features across road elements, agent interactions, and time steps. We evaluate our approach on autonomous driving datasets for both marginal and joint motion prediction, and achieve state of the art performance across two popular datasets. Through combining a scene-centric approach, agent permutation equivariant model, and a sequence masking strategy, we show that our model can unify a variety of motion prediction tasks from joint motion predictions to conditioned prediction.

DISSECT: Disentangled Simultaneous Explanations via Concept Traversals

- Asma Ghandeharioun, Been Kim, Chun-Liang Li, Brendan Jou, Brian Eoff, Rosalind Picard
- abstract@open-review(Poster): Explaining deep learning model inferences is a promising venue for scientific understanding, improving safety, uncovering hidden biases, evaluating fairness, and beyond, as argued by many scholars. One of the principal benefits of counterfactual explanations is allowing users to explore "what-if" scenarios through what does not and cannot exist in the data, a quality that many other forms of explanation such as heatmaps and influence functions are inherently incapable of doing. However, most previous work on generative explainability cannot disentangle important concepts effectively, produces unrealistic examples, or fails to retain relevant information. We propose a novel approach, DISSECT, that jointly trains a generator, a discriminator, and a concept disentangler to overcome such challenges using little supervision. DISSECT generates Concept Traversals (CTs), defined as a sequence of generated examples with increasing degrees of concepts that influence a classifier's decision. By training a generative model from a classifier's signal, DISSECT offers a way to discover a classifier's inherent "notion" of distinct concepts automatically rather than rely on user-predefined concepts. We show that DISSECT produces CTs that (1) disentangle several concepts, (2) are influential to a classifier's decision and are coupled to its reasoning due to joint training (3), are realistic, (4) preserve relevant information, and (5) are stable across similar inputs. We validate DISSECT on several challenging synthetic and realistic datasets where previous methods fall short of satisfying desirable criteria for interpretability and show that it performs consistently well. Finally, we present experiments showing applications of DISSECT for detecting potential biases of a classifier and identifying spurious artifacts that impact predictions.

Heteroscedastic Temporal Variational Autoencoder For Irregularly Sampled Time Series

- Satya Narayan Shukla, Benjamin Marlin
- abstract@open-review(Poster): Irregularly sampled time series commonly occur in several domains where they present a significant challenge to standard deep learning models. In this paper, we propose a new deep learning framework for probabilistic interpolation of irregularly sampled time series that we call the Heteroscedastic Temporal Variational Autoencoder (HeTVAE). HeTVAE includes a novel input layer to encode information about input observation sparsity, a temporal VAE architecture to propagate uncertainty due to input sparsity, and a heteroscedastic output layer to enable variable uncertainty in the output interpolations. Our results show that the proposed architecture is better able to reflect variable uncertainty through time due to sparse and irregular sampling than a range of baseline and traditional models, as well as recently proposed deep latent variable models that use homoscedastic output layers.

A Neural Tangent Kernel Perspective of Infinite Tree Ensembles

- Ryuichi Kanoh, Mahito Sugiyama
- abstract@open-review(Poster): In practical situations, the tree ensemble is one of the most popular models along with neural networks. A soft tree is a variant of a decision tree. Instead of using a greedy method for searching splitting rules, the soft tree is trained using a gradient method in which the entire splitting operation is formulated in a differentiable form. Although ensembles of such soft trees have been used increasingly in recent years, little theoretical work has been done to understand their behavior. By considering an ensemble of infinite soft trees, this paper introduces and studies the Tree Neural Tangent Kernel (TNTK), which provides new insights into the behavior of the infinite ensemble of soft trees. Using the TNTK, we theoretically identify several non-trivial properties, such as global convergence of the training, the equivalence of the oblivious tree structure, and the degeneracy of the TNTK induced by the deepening of the trees.

AlphaZero-based Proof Cost Network to Aid Game Solving

- Ti-Rong Wu, Chung-Chin Shih, Ting Han Wei, Meng-Yu Tsai, Wei-Yuan Hsu, I-Chen Wu
- abstract@open-review(Poster): The AlphaZero algorithm learns and plays games without hand-crafted expert knowledge. However, since its objective is to play well, we hypothesize that a better objective can be defined for the related but separate task of solving games. This paper proposes a novel approach to solving problems by modifying the training target of the AlphaZero algorithm, such that it prioritizes solving the game quickly, rather than winning. We train a Proof Cost Network (PCN), where proof cost is a heuristic that estimates the amount of work required to solve problems. This matches the general concept of the so-called proof number from proof number search, which has been shown to be well-suited for game solving. We propose two specific training targets. The first finds the shortest path to a solution, while the second estimates the proof cost. We conduct experiments on solving 15x15 Gomoku and 9x9 Killall-Go problems with both MCTS-based and FDFPN solvers. Comparisons between using AlphaZero networks and PCN as heuristics show that PCN can solve more problems.

Bayesian Framework for Gradient Leakage

- Mislav Balunovic, Dimitar Iliev Dimitrov, Robin Staab, Martin Vechev
- abstract@open-review(Poster): Federated learning is an established method for training machine learning models without sharing training data. However, recent work has shown that it cannot guarantee data privacy as shared gradients can still leak sensitive information. To formalize the problem of gradient leakage, we propose a theoretical framework that enables, for the first time, analysis of the Bayes optimal adversary phrased as an optimization problem. We demonstrate that existing leakage attacks can be seen as approximations of this optimal adversary with different assumptions on the probability distributions of the input data and gradients. Our experiments confirm the effectiveness of the Bayes optimal adversary when it has knowledge of the underlying distribution. Further, our experimental evaluation shows that several existing heuristic defenses are not effective against stronger attacks, especially early in the training process. Thus, our findings indicate that the construction of more effective defenses and their evaluation remains an open problem.

Universalizing Weak Supervision

- Changho Shin, Winfred Li, Harit Vishwakarma, Nicholas Carl Roberts, Frederic Sala
- abstract@open-review(Poster): Weak supervision (WS) frameworks are a popular way to bypass hand-labeling large datasets for training data-hungry models. These approaches synthesize multiple noisy but cheaply-acquired estimates of labels into a set of high-quality pseudo-labels for downstream training. However, the synthesis technique is specific to a particular kind of label, such as binary labels or sequences, and each new label type requires manually designing a new synthesis algorithm. Instead, we propose a universal technique that enables weak supervision over any label type while still offering desirable properties, including practical flexibility, computational efficiency, and theoretical guarantees. We apply this technique to important problems previously not tackled by WS frameworks including learning to rank, regression, and learning in hyperbolic space. Theoretically, our synthesis approach produces a consistent estimators for learning some challenging but important generalizations of the exponential family model. Experimentally, we validate our framework and show improvement over baselines in diverse settings including real-world learning-to-rank and regression problems along with learning on hyperbolic manifolds.

Maximum n-times Coverage for Vaccine Design

- Ge Liu, Alexander Dimitrakakis, Brandon Carter, David Gifford
- abstract@open-review(Poster): We introduce the maximum n-times coverage problem that selects k overlays to maximize the summed coverage of weighted elements, where each element must be covered at least n times. We also define the min-cost n-times coverage problem where the objective is to select the minimum set of overlays such that the sum of the weights of elements that are covered at least n times is at least τ. Maximum n-times coverage is a generalization of the multi-set multi-cover problem, is NP-complete, and is not submodular. We introduce two new practical solutions for n-times coverage based on integer linear programming and sequential greedy optimization. We show that maximum n-times coverage is a natural way to frame peptide vaccine design, and find that it produces a pan-strain COVID-19 vaccine design that is superior to 29 other published designs in predicted population coverage and the expected number of peptides displayed by each individual's HLA molecules.

KL Guided Domain Adaptation

- A. Tuan Nguyen, Toan Tran, Yarin Gal, Philip Torr, Atilim Gunes Baydin
- abstract@open-review(Poster): Domain adaptation is an important problem and often needed for real-world applications. In this problem, instead of i.i.d. training and testing datapoints, we assume that the source (training) data and the target (testing) data have different distributions. With that setting, the empirical risk minimization training procedure often does not perform well, since it does not account for the change in the distribution. A common approach in the domain adaptation literature is to learn a representation of the input that has the same (marginal) distribution over the source and the target domain. However, these approaches often require additional networks and/or optimizing an adversarial (minimax) objective, which can be very expensive or unstable in practice. To improve upon these marginal alignment techniques, in this paper, we first derive a generalization bound for the target loss based on the training loss and the reverse Kullback-Leibler (KL) divergence between the source and the target representation distributions. Based on this bound, we derive an algorithm that minimizes the KL term to obtain a better generalization to the target domain. We show that with a probabilistic representation network, the KL term can be estimated efficiently via minibatch samples without any additional network or a minimax objective. This leads to a theoretically sound alignment method which is also very efficient and stable in practice. Experimental results also suggest that our method outperforms other representation-alignment approaches.

From Stars to Subgraphs: Uplifting Any GNN with Local Structure Awareness

- Lingxiao Zhao, Wei Jin, Leman Akoglu, Neil Shah
- abstract@open-review(Poster): Message Passing Neural Networks (MPNNs) are a common type of Graph Neural Network (GNN), in which each nodeâ€™s representation is computed recursively by aggregating representations (â€œmessagesâ€) from its immediate neighbors akin to a star-shaped pattern. MPNNs are appealing for being efficient and scalable, however their expressiveness is upper-bounded by the 1st-order Weisfeiler-Lehman isomorphism test (1-WL). In response, prior works propose highly expressive models at the cost of scalability and sometimes generalization performance.

Our work stands between these two regimes: we introduce a general framework to uplift any MPNN to be more expressive, with limited scalability overhead and greatly improved practical performance. We achieve this by extending local aggregation in MPNNs from star patterns to general subgraph patterns (e.g., k-egonets): in our framework, each node representation is computed as the encoding of a surrounding induced subgraph rather than encoding of immediate neighbors only (i.e. a star). We choose the subgraph encoder to be a GNN (mainly MPNNs, considering scalability) to design a general framework that serves as a wrapper to uplift any GNN. We call our proposed method GNN-AK (GNN As Kernel), as the framework resembles a convolutional neural network by replacing the kernel with GNNs. Theoretically, we show that our framework is strictly more powerful than 1&2-WL, and is not less powerful than 3-WL. We also design subgraph sampling strategies which greatly reduce memory footprint and improve speed while maintaining performance. Our method sets new state-of-the-art performance by large margins for several well-known graph ML tasks; specifically, 0.08 MAE on ZINC, 74.79% and 86.887% accuracy on CIFAR10 and PATTERN respectively.

NETWORK INSENSITIVITY TO PARAMETER NOISE VIA PARAMETER ATTACK DURING TRAINING

- Julian BÃ¼chel, Fynn Firouz Faber, Dylan Richard Muir
- abstract@open-review(Poster): Neuromorphic neural network processors, in the form of compute-in-memory crossbar arrays of memristors, or in the form of subthreshold analog and mixed-signal ASICs, promise enormous advantages in compute density and energy efficiency for NN-based ML tasks. However, these technologies are prone to computational non-idealities, due to process variation and intrinsic device physics. This degrades the task performance of networks deployed to the processor, by introducing parameter noise into the deployed model. While it is possible to calibrate each device, or train networks individually for each processor, these approaches are expensive and impractical for commercial deployment. Alternative methods are therefore needed to train networks that are inherently robust against parameter variation, as a consequence of network architecture and parameters. We present a new network training algorithm that attacks network parameters during training, and promotes robust performance during inference in the face of random parameter variation. Our approach introduces a loss regularization term that penalizes the susceptibility of a network to weight perturbation. We compare against previous approaches for producing parameter insensitivity such as dropout, weight smoothing and introducing parameter noise during training. We show that our approach produces models that are more robust to random mismatch-induced parameter variation as well as to targeted parameter variation. Our approach finds minima in flatter locations in the weight-loss landscape compared with other approaches, highlighting that the networks found by our technique are less sensitive to parameter perturbation. Our work provides an approach to deploy neural network architectures to inference devices that suffer from computational non-idealities, with minimal loss of performance. This method will enable deployment at scale to novel energy-efficient computational substrates, promoting cheaper and more prevalent edge inference.

Gradient Importance Learning for Incomplete Observations

- Qitong Gao, Dong Wang, Joshua David Amason, Siyang Yuan, Chenyang Tao, Ricardo Henao, Majda Hadziahmetovic, Lawrence Carin, Miroslav Pajic
- abstract@open-review(Poster): Though recent works have developed methods that can generate estimates (or imputations) of the missing entries in a dataset to facilitate downstream analysis, most depend on assumptions that may not align with real-world applications and could suffer from poor performance in subsequent tasks such as classification. This is particularly true if the data have large missingness rates or a small sample size. More importantly, the imputation error could be propagated into the prediction step that follows, which may constrain the capabilities of the prediction model. In this work, we introduce the gradient importance learning (GIL) method to train multilayer perceptrons (MLPs) and long short-term memories (LSTMs) to directly perform inference from inputs containing missing values without imputation. Specifically, we employ reinforcement learning (RL) to adjust the gradients used to train these models via back-propagation. This allows the model to exploit the underlying information behind missingness patterns. We test the approach on real-world time-series (i.e., MIMIC-III), tabular data obtained from an eye clinic, and a standard dataset (i.e., MNIST), where our imputation-free predictions outperform the traditional two-step imputation-based predictions using state-of-the-art imputation methods.

Do Users Benefit From Interpretable Vision? A User Study, Baseline, And Dataset

- Leon Sixt, Martin Schuessler, Oana-Iuliana Popescu, Philipp WeiÃŸ, Tim Landgraf
- abstract@open-review(Poster): A variety of methods exist to explain image classification models. However, whether they provide any benefit to users over simply comparing various inputs and the modelâ€™s respective predictions remains unclear. We conducted a user study (N=240) to test how such a baseline explanation technique performs against concept-based and counterfactual explanations. To this end, we contribute a synthetic dataset generator capable of biasing individual attributes and quantifying their relevance to the model. In a study, we assess if participants can identify the relevant set of attributes compared to the ground-truth. Our results show that the baseline outperformed concept-based explanations. Counterfactual explanations from an invertible neural network performed similarly as the baseline. Still, they allowed users to identify some attributes more accurately. Our results highlight the importance of measuring how well users can reason about biases of a model, rather than solely relying on technical evaluations or proxy tasks. We open-source our study and dataset so it can serve as a blue-print for future studies.

Understanding the Variance Collapse of SVGD in High Dimensions

- Jimmy Ba, Murat A Erdogdu, Marzyeh Ghassemi, Shengyang Sun, Taiji Suzuki, Denny Wu, Tianzong Zhang
- abstract@open-review(Poster): Stein variational gradient descent (SVGD) is a deterministic inference algorithm that evolves a set of particles to fit a target distribution. Despite its computational efficiency, SVGD often underestimates the variance of the target distribution in high dimensions. In this work we attempt to explain the variance collapse in SVGD. On the qualitative side, we compare the SVGD update with gradient descent on the maximum mean discrepancy (MMD) objective; we observe that the variance collapse phenomenon relates to the bias from deterministic updates present in the "driving force" of SVGD, and empirically verify that removal of such bias leads to more accurate variance estimation. On the quantitative side, we demonstrate that the variance collapse of SVGD can be accurately predicted in the proportional asymptotic limit, i.e., when the number of particles n and dimensions d diverge at the same rate. In particular, for learning high-dimensional isotropic Gaussians, we derive the exact equilibrium variance for both SVGD and MMD-descent under certain near-orthogonality assumption on the converged particles, and confirm that SVGD suffers from the "curse of dimensionality".

Generalisation in Lifelong Reinforcement Learning through Logical Composition

- Geraud Nangue Tasse, Steven James, Benjamin Rosman
- abstract@open-review(Poster): We leverage logical composition in reinforcement learning to create a framework that enables an agent to autonomously determine whether a new task can be immediately solved using its existing abilities, or whether a task-specific skill should be learned. In the latter case, the proposed algorithm also enables the agent to learn the new task faster by generating an estimate of the optimal policy. Importantly, we provide two main theoretical results: we bound the performance of the transferred policy on a new task, and we give bounds on the necessary and sufficient number of tasks that need to be learned throughout an agent's lifetime to generalise over a distribution. We verify our approach in a series of experiments, where we perform transfer learning both after learning a set of base tasks, and after learning an arbitrary set of tasks. We also demonstrate that, as a side effect of our transfer learning approach, an agent can produce an interpretable Boolean expression of its understanding of the current task. Finally, we demonstrate our approach in the full lifelong setting where an agent receives tasks from an unknown distribution. Starting from scratch, an agent is able to quickly generalise over the task distribution after learning only a few tasks, which are sub-logarithmic in the size of the task space.

PolyLoss: A Polynomial Expansion Perspective of Classification Loss Functions

- Zhaoqi Leng, Mingxing Tan, Chenxi Liu, Ekin Dogus Cubuk, Jay Shi, Shuyang Cheng, Dragomir Anguelov

- abstract@open-review(Poster): Cross-entropy loss and focal loss are the most common choices when training deep neural networks for classification problems. Generally speaking, however, a good loss function can take on much more flexible forms, and should be tailored for different tasks and datasets. Motivated by how functions can be approximated via Taylor expansion, we propose a simple framework, named PolyLoss, to view and design loss functions as a linear combination of polynomial functions. Our PolyLoss allows the importance of different polynomial bases to be easily adjusted depending on the targeting tasks and datasets, while naturally subsuming the aforementioned cross-entropy loss and focal loss as special cases. Extensive experimental results show that the optimal choice within the PolyLoss is indeed dependent on the task and dataset. Simply by introducing one extra hyperparameter and adding one line of code, our Poly-1 formulation outperforms the cross-entropy loss and focal loss on 2D image classification, instance segmentation, object detection, and 3D object detection tasks, sometimes by a large margin.

Improving Non-Autoregressive Translation Models Without Distillation

- Xiao Shi Huang, Felipe Perez, Maksims Volkovs
- abstract@open-review(Poster): Transformer-based autoregressive (AR) machine translation models have achieved significant performance improvements, nearing human-level accuracy on some languages. The AR framework translates one token at a time which can be time consuming, especially for long sequences. To accelerate inference, recent work has been exploring non-autoregressive (NAR) approaches that translate blocks of tokens in parallel. Despite significant progress, leading NAR models still lag behind their AR counterparts, and only become competitive when trained with distillation. In this paper we investigate possible reasons behind this performance gap, namely, the indistinguishability of tokens, and mismatch between training and inference. We then propose the Conditional Masked Language Model with Correction (CMLMC) that addresses these problems. Empirically, we show that CMLMC achieves state-of-the-art NAR performance when trained on raw data without distillation and approaches AR performance on multiple datasets. Full code for this work will be released at the time of publication.

A Theory of Tournament Representations

- Arun Rajkumar, Vishnu Veerathu, Abdul Bakey Mir
- abstract@open-review(Poster): Real-world tournaments are almost always intransitive. Recent works have noted that parametric models which assume d dimensional node representations can effectively model intransitive tournaments. However, nothing is known about the structure of the class of tournaments that arise out of any fixed d dimensional representations. In this work, we develop a novel theory for understanding parametric tournament representations. Our first contribution is to structurally characterize the class of tournaments that arise out of d dimensional representations. We do this by showing that these tournament classes have forbidden configurations that must necessarily be a union of flip classes, a novel way to partition the set of all tournaments. We further characterize rank 2 tournaments completely by showing that the associated forbidden flip class contains just 2 tournaments. Specifically, we show that the rank 2 tournaments are equivalent to locally transitive tournaments. This insight allows us to show that the minimum feedback arc set problem on this tournament class can be solved using the standard Quicksort procedure. We also exhibit specific forbidden configurations for rank 4 tournaments. For a general rank d tournament class, we show that the flip class associated with a coned-doubly regular tournament of size $\mathcal{O}(\sqrt{d})$ must be a forbidden configuration. To answer a dual question, using a celebrated result of Froster, we show a lower bound of $\Theta(\sqrt{n})$ on the minimum dimension needed to represent all tournaments on n nodes. For any given tournament, we show a novel upper bound on the smallest representation dimension that depends on the least size of the number of unique nodes in any feedback arc set of the flip class associated with a tournament. We show how our results also shed light on the upper bound of sign-rank of matrices.

Convergent and Efficient Deep Q Learning Algorithm

- Zhikang T. Wang, Masahito Ueda
- abstract@open-review(Poster): Despite the empirical success of the deep Q network (DQN) reinforcement learning algorithm and its variants, DQN is still not well understood and it does not guarantee convergence. In this work, we show that DQN can indeed diverge and cease to operate in realistic settings. Although there exist gradient-based convergent methods, we show that they actually have inherent problems in learning dynamics which cause them to fail even for simple tasks. To overcome these problems, we propose a convergent DQN algorithm (C-DQN) that is guaranteed to converge and can work with large discount factors (0.9998). It learns robustly in difficult settings and can learn several difficult games in the Atari 2600 benchmark that DQN fails to solve.

Trigger Hunting with a Topological Prior for Trojan Detection

- Xiaoling Hu, Xiao Lin, Michael Cogswell, Yi Yao, Susmit Jha, Chao Chen
- abstract@open-review(Poster): Despite their success and popularity, deep neural networks (DNNs) are vulnerable when facing backdoor attacks. This impedes their wider adoption, especially in mission critical applications. This paper tackles the problem of Trojan detection, namely, identifying Trojaned models â€“ models trained with poisoned data. One popular approach is reverse engineering, i.e., recovering the triggers on a clean image by manipulating the modelâ€™s prediction. One major challenge of reverse engineering approach is the enormous search space of triggers. To this end, we propose innovative priors such as diversity and topological simplicity to not only increase the chances of finding the appropriate triggers but also improve the quality of the found triggers. Moreover, by encouraging a diverse set of trigger candidates, our method can perform effectively in cases with unknown target labels. We demonstrate that these priors can significantly improve the quality of the recovered triggers, resulting in substantially improved Trojan detection accuracy as validated on both synthetic and publicly available TrojAI benchmarks.

Who Is the Strongest Enemy? Towards Optimal and Efficient Evasion Attacks in Deep RL

- Yanchao Sun, Ruijie Zheng, Yongyuan Liang, Furong Huang
- abstract@open-review(Poster): Evaluating the worst-case performance of a reinforcement learning (RL) agent under the strongest/optimal adversarial perturbations on state observations (within some constraints) is crucial for understanding the robustness of RL agents. However, finding the optimal adversary is challenging, in terms of both whether we can find the optimal attack and how efficiently we can find it. Existing works on adversarial RL either use heuristics-based methods that may not find the strongest adversary, or directly train an RL-based adversary by treating the agent as a part of the environment, which can find the optimal adversary but may become intractable in a large state space. This paper introduces a novel attacking method to find the optimal attacks through collaboration between a designed function named "actor" and an RL-based learner named "director'". The actor crafts state perturbations for a given policy perturbation direction, and the director learns to propose the best policy perturbation directions. Our proposed algorithm, PA-AD, is theoretically optimal and significantly more efficient than prior RL-based works in environments with large state spaces. Empirical results show that our proposed PA-AD universally outperforms state-of-the-art attacking methods in various Atari and MuJoCo environments. By applying PA-AD to adversarial training, we achieve state-of-the-art empirical robustness in multiple tasks under strong adversaries.

Chunked Autoregressive GAN for Conditional Waveform Synthesis

- Max Morrison, Rithesh Kumar, Kundan Kumar, Prem Seetharaman, Aaron Courville, Yoshua Bengio
- abstract@open-review(Poster): Conditional waveform synthesis models learn a distribution of audio waveforms given conditioning such as text, mel-spectrograms, or MIDI. These systems employ deep generative models that model the waveform via either sequential (autoregressive) or parallel (non-autoregressive) sampling. Generative adversarial networks (GANs) have become a common choice for non-autoregressive waveform synthesis. However, state-of-the-art GAN-based models produce artifacts when performing mel-spectrogram inversion. In this paper, we demonstrate that these artifacts correspond with an inability for the generator to learn accurate pitch and periodicity. We show that simple pitch and periodicity conditioning is insufficient

for reducing this error relative to using autoregression. We discuss the inductive bias that autoregression provides for learning the relationship between instantaneous frequency and phase, and show that this inductive bias holds even when autoregressively sampling large chunks of the waveform during each forward pass. Relative to prior state-of-the-art GAN-based models, our proposed model, Chunked Autoregressive GAN (CARGAN) reduces pitch error by 40-60%, reduces training time by 58%, maintains a fast inference speed suitable for real-time or interactive applications, and maintains or improves subjective quality.

COPA: Certifying Robust Policies for Offline Reinforcement Learning against Poisoning Attacks

- Fan Wu, Linyi Li, Huan Zhang, Bhavya Kailkhura, Krishnaram Kenthapadi, Ding Zhao, Bo Li
- abstract@open-review(Poster): As reinforcement learning (RL) has achieved near human-level performance in a variety of tasks, its robustness has raised great attention. While a vast body of research has explored test-time (evasion) attacks in RL and corresponding defenses, its robustness against training-time (poisoning) attacks remains largely unanswered. In this work, we focus on certifying the robustness of ofï¬‚ine RL in the presence of poisoning attacks, where a subset of training trajectories could be arbitrarily manipulated. We propose the ï¬ rst certiï¬ cation framework, COPA, to certify the number of poisoning trajectories that can be tolerated regarding different certiï¬ cation criteria. Given the complex structure of RL, we propose two certiï¬ cation criteria: per-state action stability and cumulative reward bound. To further improve the certiï¬ cation, we propose new partition and aggregation protocols to train robust policies. We further prove that some of the proposed certiï¬ cation methods are theoretically tight and some are NP-Complete problems. We leverage COPA to certify three RL environments trained with different algorithms and conclude: (1) The proposed robust aggregation protocols such as temporal aggregation can signiï¬ cantly improve the certiï¬ cations; (2) Our certiï¬ cations for both per-state action stability and cumulative reward bound are efï¬ cient and tight; (3) The certiï¬ cation for different training algorithms and environments are different, implying their intrinsic robustness properties. All experimental results are available at https://copa-leaderboard.github.io.

ExT5: Towards Extreme Multi-Task Scaling for Transfer Learning

- Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Gupta, Kai Hui, Sebastian Ruder, Donald Metzler
- abstract@open-review(Poster): Despite the recent success of multi-task learning and transfer learning for natural language processing (NLP), few works have systematically studied the effect of scaling up the number of tasks during pre-training. Towards this goal, this paper introduces ExMix (Extreme Mixture): a massive collection of 107 supervised NLP tasks across diverse domains and task-families. Using ExMix, we study the effect of multi-task pre-training at the largest scale to date, and analyze co-training transfer amongst common families of tasks. Through this analysis, we show that manually curating an ideal set of tasks for multi-task pre-training is not straightforward, and that multi-task scaling can vastly improve models on its own. Finally, we propose ExT5: a model pre-trained using a multi-task objective of self-supervised span denoising and supervised ExMix. Via extensive experiments, we show that ExT5 outperforms strong T5 baselines on SuperGLUE, GEM, Rainbow, Closed-Book QA tasks, and several tasks outside of ExMix. ExT5 also significantly improves sample efficiency while pre-training.

Provable Adaptation across Multiway Domains via Representation Learning

- Zhili Feng, Shaobo Han, Simon Shaolei Du
- abstract@open-review(Poster): This paper studies zero-shot domain adaptation where each domain is indexed on a multi-dimensional array, and we only have data from a small subset of domains. Our goal is to produce predictors that perform well on \emph{unseen} domains. We propose a model which consists of a domain-invariant latent representation layer and a domain-specific linear prediction layer with a low-rank tensor structure. Theoretically, we present explicit sample complexity bounds to characterize the prediction error on unseen domains in terms of the number of domains with training data and the number of data per domain. To our knowledge, this is the first finite-sample guarantee for zero-shot domain adaptation. In addition, we provide experiments on two-way MNIST and four-way fiber sensing datasets to demonstrate the effectiveness of our proposed model.

Efficient Token Mixing for Transformers via Adaptive Fourier Neural Operators

- John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, Bryan Catanzaro
- abstract@open-review(Poster): Vision transformers have delivered tremendous success in representation learning. This is primarily due to effective token mixing through self attention. However, this scales quadratically with the number of pixels, which becomes infeasible for high-resolution inputs. To cope with this challenge, we propose Adaptive Fourier Neural Operator (AFNO) as an efficient token mixer that learns to mix in the Fourier domain. AFNO is based on a principled foundation of operator learning which allows us to frame token mixing as a continuous global convolution without any dependence on the input resolution. This principle was previously used to design FNO, which solves global convolution efficiently in the Fourier domain and has shown promise in learning challenging PDEs. To handle challenges in visual representation learning such as discontinuities in images and high resolution inputs, we propose principled architectural modifications to FNO which results in memory and computational efficiency. This includes imposing a block-diagonal structure on the channel mixing weights, adaptively sharing weights across tokens, and sparsifying the frequency modes via soft-thresholding and shrinkage. The resulting model is highly parallel with a quasi-linear complexity and has linear memory in the sequence size. AFNO outperforms self-attention mechanisms for few-shot segmentation in terms of both efficiency and accuracy. For Cityscapes segmentation with the Segformer-B3 backbone, AFNO can handle a sequence size of 65k and outperforms other efficient self-attention mechanisms.

Sample Selection with Uncertainty of Losses for Learning with Noisy Labels

- Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Jun Yu, Gang Niu, Masashi Sugiyama
- abstract@open-review(Poster): In learning with noisy labels, the sample selection approach is very popular, which regards small-loss data as correctly labeled data during training. However, losses are generated on-the-ï¬‚y based on the model being trained with noisy labels, and thus large-loss data are likely but not certain to be incorrect. There are actually two possibilities of a large-loss data point: (a) it is mislabeled, and then its loss decreases slower than other data, since deep neural networks learn patterns ï¬ rst; (b) it belongs to an underrepresented group of data and has not been selected yet. In this paper, we incorporate the uncertainty of losses by adopting interval estimation instead of point estimation of losses, where lower bounds of the conï¬ dence intervals of losses derived from distribution-free concentration inequalities, but not losses themselves, are used for sample selection. In this way, we also give large-loss but less selected data a try; then, we can better distinguish between the cases (a) and (b) by seeing if the losses effectively decrease with the uncertainty after the try. As a result, we can better explore underrepresented data that are correctly labeled but seem to be mislabeled at ï¬ rst glance. Experiments demonstrate that the proposed method is superior to baselines and robust to a broad range of label noise types.

Data-Driven Offline Optimization for Architecting Hardware Accelerators

- Aviral Kumar, Amir Yazdanbakhsh, Milad Hashemi, Kevin Swersky, Sergey Levine
- abstract@open-review(Poster): To attain higher efficiency, the industry has gradually reformed towards application-specific hardware accelerators. While such a paradigm shift is already starting to show promising results, designers need to spend considerable manual effort and perform large number of time-consuming simulations to find accelerators that can accelerate multiple target applications while obeying design constraints. Moreover, such a simulation-driven approach must be re-run from scratch every time the set of target applications or design constraints change. An alternative paradigm is to use a data-driven, offline approach that utilizes logged simulation data, to architect hardware accelerators, without needing any form of simulations. Such an

approach not only alleviates the need to run time-consuming simulation, but also enables data reuse and applies even when set of target applications changes. In this paper, we develop such a data-driven offline optimization method for designing hardware accelerators, dubbed PRIME, that enjoys all of these properties. Our approach learns a conservative, robust estimate of the desired cost function, utilizes infeasible points and optimizes the design against this estimate without any additional simulator queries during optimization. PRIME architects accelerators---tailored towards both single- and multi-applications---improving performance upon stat-of-the-art simulation-driven methods by about 1.54x and 1.20x, while considerably reducing the required total simulation time by 93% and 99%, respectively. In addition, PRIME also architects effective accelerators for unseen applications in a zero-shot setting, outperforming simulation-based methods by 1.26x.

Multi-Agent MDP Homomorphic Networks

- Elise van der Pol, Herke van Hoof, Frans A Oliehoek, Max Welling
- abstract@open-review(Poster): This paper introduces Multi-Agent MDP Homomorphic Networks, a class of networks that allows distributed execution using only local information, yet is able to share experience between global symmetries in the joint state-action space of cooperative multi-agent systems. In cooperative multi-agent systems, complex symmetries arise between different configurations of the agents and their local observations. For example, consider a group of agents navigating: rotating the state globally results in a permutation of the optimal joint policy. Existing work on symmetries in single agent reinforcement learning can only be generalized to the fully centralized setting, because such approaches rely on the global symmetry in the full state-action spaces, and these can result in correspondences across agents. To encode such symmetries while still allowing distributed execution we propose a factorization that decomposes global symmetries into local transformations. Our proposed factorization allows for distributing the computation that enforces global symmetries over local agents and local interactions. We introduce a multi-agent equivariant policy network based on this factorization. We show empirically on symmetric multi-agent problems that globally symmetric distributable policies improve data efficiency compared to non-equivariant baselines.

Geometry-Consistent Neural Shape Representation with Implicit Displacement Fields

- Wang Yifan, Lukas Rahmann, Olga Sorkine-hornung
- abstract@open-review(Poster): We present implicit displacement fields, a novel representation for detailed 3D geometry. Inspired by a classic surface deformation technique, displacement mapping, our method represents a complex surface as a smooth base surface plus a displacement along the base's normal directions, resulting in a frequency-based shape decomposition, where the high-frequency signal is constrained geometrically by the low-frequency signal. Importantly, this disentanglement is unsupervised thanks to a tailored architectural design that has an innate frequency hierarchy by construction. We explore implicit displacement field surface reconstruction and detail transfer and demonstrate superior representational power, training stability, and generalizability.

Modeling Label Space Interactions in Multi-label Classification using Box Embeddings

- Dhruvesh Patel, Pavitra Dangati, Jay-Yoon Lee, Michael Boratko, Andrew McCallum
- abstract@open-review(Poster): Multi-label classification is a challenging structured prediction task in which a set of output class labels are predicted for each input. Real-world datasets often have natural or latent taxonomic relationships between labels, making it desirable for models to employ label representations capable of capturing such taxonomies. Most existing multi-label classification methods do not do so, resulting in label predictions that are inconsistent with the taxonomic constraints, thus failing to accurately represent the fundamentals of problem setting. In this work, we introduce the multi-label box model (MBM), a multi-label classification method that combines the encoding power of neural networks with the inductive bias and probabilistic semantics of box embeddings (Vilnis, et al 2018). Box embeddings can be understood as trainable Venn-diagrams based on hyper-rectangles. Representing labels by boxes rather than vectors, MBM is able to capture taxonomic relations among labels. Furthermore, since box embeddings allow these relations to be learned by stochastic gradient descent from data, and to be read as calibrated conditional probabilities, our model is endowed with a high degree of interpretability. This interpretability also facilitates the injection of partial information about label-label relationships into model training, to further improve its consistency. We provide theoretical grounding for our method and show experimentally the model's ability to learn the true latent taxonomic structure from data. Through extensive empirical evaluations on both small and large-scale multi-label classification datasets, we show that BBM can significantly improve taxonomic consistency while preserving or surpassing the state-of-the-art predictive performance.

It Takes Two to Tango: Mixup for Deep Metric Learning

- Shashanka Venkataramanan, Bill Psomas, Ewa Kijak, laurent amsaleg, Konstantinos Karantzalos, Yannis Avrithis
- abstract@open-review(Poster): Metric learning involves learning a discriminative representation such that embeddings of similar classes are encouraged to be close, while embeddings of dissimilar classes are pushed far apart. State-of-the-art methods focus mostly on sophisticated loss functions or mining strategies. On the one hand, metric learning losses consider two or more examples at a time. On the other hand, modern data augmentation methods for classification consider two or more examples at a time. The combination of the two ideas is under-studied.

In this work, we aim to bridge this gap and improve representations using mixup, which is a powerful data augmentation approach interpolating two or more examples and corresponding target labels at a time. This task is challenging because, unlike classification, the loss functions used in metric learning are not additive over examples, so the idea of interpolating target labels is not straightforward. To the best of our knowledge, we are the first to investigate mixing both examples and target labels for deep metric learning. We develop a generalized formulation that encompasses existing metric learning loss functions and modify it to accommodate for mixup, introducing Metric Mix, or Metrix. We also introduce a new metric---utilization---to demonstrate that by mixing examples during training, we are exploring areas of the embedding space beyond the training classes, thereby improving representations. To validate the effect of improved representations, we show that mixing inputs, intermediate representations or embeddings along with target labels significantly outperforms state-of-the-art metric learning methods on four benchmark deep metric learning datasets.

Data Efficient Language-Supervised Zero-Shot Recognition with Optimal Transport Distillation

- Bichen Wu, Ruizhe Cheng, Peizhao Zhang, Tianren Gao, Joseph E. Gonzalez, Peter Vajda
- abstract@open-review(Poster): Traditional computer vision models are trained to predict a fixed set of predefined categories. Recently, natural language has been shown to be a broader and richer source of supervision that provides finer descriptions to visual concepts than supervised "gold" labels. Previous works, such as CLIP, use InfoNCE loss to train a model to predict the pairing between images and text captions. CLIP, however, is data hungry and requires more than 400M image-text pairs for training. The inefficiency can be \textit{partially} attributed to the fact that the image-text pairs are noisy. To address this, we propose OTTER (Optimal TransporT distillation for Efficient zero-shot Recognition), which uses online entropic optimal transport to find a soft image-text match as labels for contrastive learning. Based on pretrained image and text encoders, models trained with OTTER achieve strong performance with only 3M image text pairs. Compared with InfoNCE loss, label smoothing, and knowledge distillation, OTTER consistently outperforms these baselines in zero-shot evaluation on Google Open Images (19,958 classes) and multi-labeled ImageNet 10K (10032 classes) from Tencent ML-Images. Over 42 evaluations on 7 different dataset/architecture settings x 6 metrics, OTTER outperforms (32) or ties (2) all baselines in 34 of them. Our source code is open sourced at https://github.com/facebookresearch/OTTER.

A Statistical Framework for Efficient Out of Distribution Detection in Deep Neural Networks

- Matan Haroush, Tzviel Frostig, Ruth Heller, Daniel Soudry

- abstract@open-review(Poster): Background. Commonly, Deep Neural Networks (DNNs) generalize well on samples drawn from a distribution similar to that of the training set. However, DNNs' predictions are brittle and unreliable when the test samples are drawn from a dissimilar distribution. This is a major concern for deployment in real-world applications, where such behavior may come at a considerable cost, such as industrial production lines, autonomous vehicles, or healthcare applications.

Contributions. We frame Out Of Distribution (OOD) detection in DNNs as a statistical hypothesis testing problem. Tests generated within our proposed framework combine evidence from the entire network. Unlike previous OOD detection heuristics, this framework returns a p-value for each test sample. It is guaranteed to maintain the Type I Error (T1E - incorrectly predicting OOD for an actual in-distribution sample) for test data. Moreover, this allows to combine several detectors while maintaining the T1E.

Building on this framework, we suggest a novel OOD procedure based on low-order statistics. Our method achieves comparable or better results than state-of-the-art methods on well-accepted OOD benchmarks, without retraining the network parameters or assuming prior knowledge on the test distribution --- and at a fraction of the computational cost.

FedBABU: Toward Enhanced Representation for Federated Image Classification

- Jaehoon Oh, SangMook Kim, Se-Young Yun
- abstract@open-review(Poster): Federated learning has evolved to improve a single global model under data heterogeneity (as a curse) or to develop multiple personalized models using data heterogeneity (as a blessing). However, little research has considered both directions simultaneously. In this paper, we first investigate the relationship between them by analyzing Federated Averaging at the client level and determine that a better federated global model performance does not constantly improve personalization. To elucidate the cause of this personalization performance degradation problem, we decompose the entire network into the body (extractor), which is related to universality, and the head (classifier), which is related to personalization. We then point out that this problem stems from training the head. Based on this observation, we propose a novel federated learning algorithm, coined FedBABU, which only updates the body of the model during federated training (i.e., the head is randomly initialized and never updated), and the head is fine-tuned for personalization during the evaluation process. Extensive experiments show consistent performance improvements and an efficient personalization of FedBABU. The code is available at https://github.com/jhoon-oh/FedBABU.

Should I Run Offline Reinforcement Learning or Behavioral Cloning?

- Aviral Kumar, Joey Hong, Anikait Singh, Sergey Levine
- abstract@open-review(Poster): Offline reinforcement learning (RL) algorithms can acquire effective policies by utilizing only previously collected experience, without any online interaction. While it is widely understood that offline RL is able to extract good policies even from highly suboptimal data, in practice offline RL is often used with data that resembles demonstrations. In this case, one can also use behavioral cloning (BC) algorithms, which mimic a subset of the dataset via supervised learning. It seems natural to ask: When should we prefer offline RL over BC? In this paper, our goal is to characterize environments and dataset compositions where offline RL leads to better performance than BC. In particular, we characterize the properties of environments that allow offline RL methods to perform better than BC methods even when only provided with expert data. Additionally, we show that policies trained on suboptimal data that is sufficiently noisy can attain better performance than even BC algorithms with expert data, especially on long-horizon problems. We validate our theoretical results via extensive experiments on both diagnostic and high-dimensional domains including robot manipulation, maze navigation and Atari games, when learning from a variety of data sources. We observe that modern offline RL methods trained on suboptimal, noisy data in sparse reward domains outperform cloning the expert data in several practical problems.

Learning State Representations via Retracing in Reinforcement Learning

- Changmin Yu, Dong Li, Jianye HAO, Jun Wang, Neil Burgess
- abstract@open-review(Poster): We propose learning via retracing, a novel self-supervised approach for learning the state representation (and the associated dynamics model) for reinforcement learning tasks. In addition to the predictive (reconstruction) supervision in the forward direction, we propose to include "retraced" transitions for representation/model learning, by enforcing the cycle-consistency constraint between the original and retraced states, hence improve upon the sample efficiency of learning. Moreover, learning via retracing explicitly propagates information about future transitions backward for inferring previous states, thus facilitates stronger representation learning for the downstream reinforcement learning tasks. We introduce Cycle-Consistency World Model (CCWM), a concrete model-based instantiation of learning via retracing. Additionally we propose a novel adaptive "truncation" mechanism for counteracting the negative impacts brought by "irreversible" transitions such that learning via retracing can be maximally effective. Through extensive empirical studies on visual-based continuous control benchmarks, we demonstrate that CCWM achieves state-of-the-art performance in terms of sample efficiency and asymptotic performance, whilst exhibiting behaviours that are indicative of stronger representation learning.

Open-World Semi-Supervised Learning

- Kaidi Cao, Maria Brbic, Jure Leskovec
- abstract@open-review(Poster): A fundamental limitation of applying semi-supervised learning in real-world settings is the assumption that unlabeled test data contains only classes previously encountered in the labeled training data. However, this assumption rarely holds for data in-the-wild, where instances belonging to novel classes may appear at testing time. Here, we introduce a novel open-world semi-supervised learning setting that formalizes the notion that novel classes may appear in the unlabeled test data. In this novel setting, the goal is to solve the class distribution mismatch problem between labeled and unlabeled data, where at the test time every input instance either needs to be classified into one of the existing classes or a new unseen class needs to be initialized and the instance assigned to it. To tackle this challenging problem, we propose ORCA, an end-to-end approach that assigns instances to previously seen classes or forms novel classes by grouping similar instances without assuming any prior knowledge. The key idea in ORCA is to utilize uncertainty adaptive margin to circumvent the bias towards seen classes caused by learning seen classes faster than the novel classes. In this way, ORCA gradually increases the discriminability of the model during the training and reduces the gap between intra-class variance of seen with respect to novel classes. Extensive experiments on image classification datasets and a single-cell dataset demonstrate that ORCA consistently outperforms alternative baselines, achieving 25% improvement on seen and 96% improvement on novel classes of the ImageNet dataset.

Evading Adversarial Example Detection Defenses with Orthogonal Projected Gradient Descent

- Oliver Bryniarski, Nabeel Hingun, Pedro Pachuca, Vincent Wang, Nicholas Carlini
- abstract@open-review(Poster): Evading adversarial example detection defenses requires finding adversarial examples that must simultaneously (a) be misclassified by the model and (b) be detected as non-adversarial. We find that existing attacks that attempt to satisfy multiple simultaneous constraints often over-optimize against one constraint at the cost of satisfying another. We introduce Selective Projected Gradient Descent and Orthogonal Projected Gradient Descent, improved attack techniques to generate adversarial examples that avoid this problem by orthogonalizing the gradients when running standard gradient-based attacks. We use our technique to evade four state-of-the-art detection defenses, reducing their accuracy to 0% while maintaining a 0% detection rate.

Reducing Excessive Margin to Achieve a Better Accuracy vs. Robustness Trade-off

- Rahul Rade, Seyed-Mohsen Moosavi-Dezfooli
- abstract@open-review(Poster): While adversarial training has become the de facto approach for training robust classifiers, it leads to a drop in accuracy. This has led to prior works postulating that accuracy is inherently at odds with robustness. Yet, the phenomenon remains inexplicable. In this paper, we closely examine the changes induced in the decision boundary of a deep network during adversarial training. We find that adversarial training leads to unwarranted increase in the margin along certain adversarial directions, thereby hurting accuracy. Motivated by this observation, we present a novel algorithm, called Helper-based Adversarial Training (HAT), to reduce this effect by incorporating additional wrongly labelled examples during training. Our proposed method provides a notable improvement in accuracy without compromising robustness. It achieves a better trade-off between accuracy and robustness in comparison to existing defenses. Code is available at https://github.com/imrahulr/hat.

Expressivity of Emergent Languages is a Trade-off between Contextual Complexity and Unpredictability

- Shangmin Guo, Yi Ren, Kory Wallace Mathewson, Simon Kirby, Stefano V Albrecht, Kenny Smith
- abstract@open-review(Poster): Researchers are using deep learning models to explore the emergence of language in various language games, where agents interact and develop an emergent language to solve tasks. We focus on the factors that determine the expressivity of emergent languages, which reflects the amount of information about input spaces those languages are capable of encoding. We measure the expressivity of emergent languages based on the generalisation performance across different games, and demonstrate that the expressivity of emergent languages is a trade-off between the complexity and unpredictability of the context those languages emerged from. Another contribution of this work is the discovery of message type collapse, i.e. the number of unique messages is lower than that of inputs. We also show that using the contrastive loss proposed by Chen et al. (2020) can alleviate this problem.

Fast AdvProp

- Jieru Mei, Yucheng Han, Yutong Bai, Yixiao Zhang, Yingwei Li, Xianhang Li, Alan Yuille, Cihang Xie
- abstract@open-review(Poster): Adversarial Propagation (AdvProp) is an effective way to improve recognition models, leveraging adversarial examples. Nonetheless, AdvProp suffers from the extremely slow training speed, mainly because: a) extra forward and backward passes are required for generating adversarial examples; b) both original samples and their adversarial counterparts are used for training (i.e., 2X data). In this paper, we introduce Fast AdvProp, which aggressively revamps AdvProp's costly training components, rendering the method nearly as cheap as the vanilla training. Specifically, our modifications in Fast AdvProp are guided by the hypothesis that disentangled learning with adversarial examples is the key for performance improvements, while other training recipes (e.g., paired clean and adversarial training samples, multi-step adversarial attackers) could be largely simplified.

Our empirical results show that, compared to the vanilla training baseline, Fast AdvProp is able to further model performance on a spectrum of visual benchmarks, without incurring extra training cost. Additionally, our ablations find Fast AdvProp scales better if larger models are used, is compatible with existing data augmentation methods (i.e., Mixup and CutMix), and can be easily adapted to other recognition tasks like object detection. The code is available here: https://github.com/meijieru/fast_advprop.

Triangle and Four Cycle Counting with Predictions in Graph Streams

- Justin Y Chen, Talya Eden, Piotr Indyk, Honghao Lin, Shyam Narayanan, Ronitt Rubinfeld, Sandeep Silwal, Tal Wagner, David Woodruff, Michael Zhang
- abstract@open-review(Poster): We propose data-driven one-pass streaming algorithms for estimating the number of triangles and four cycles, two fundamental problems in graph analytics that are widely studied in the graph data stream literature. Recently, Hsu et al. (2019) and Jiang et al. (2020) applied machine learning techniques in other data stream problems, using a trained oracle that can predict certain properties of the stream elements to improve on prior â€œclassicalâ€ algorithms that did not use oracles. In this paper, we explore the power of a â€œheavy edgeâ€ oracle in multiple graph edge streaming models. In the adjacency list model, we present a one-pass triangle counting algorithm improving upon the previous space upper bounds without such an oracle. In the arbitrary order model, we present algorithms for both triangle and four cycle estimation with fewer passes and the same space complexity as in previous algorithms, and we show several of these bounds are optimal. We analyze our algorithms under several noise models, showing that the algorithms perform well even when the oracle errs. Our methodology expands upon prior work on â€œclassicalâ€ streaming algorithms, as previous multi-pass and random order streaming algorithms can be seen as special cases of our algorithms, where the first pass or random order was used to implement the heavy edge oracle. Lastly, our experiments demonstrate advantages of the proposed method compared to state-of-the-art streaming algorithms.

Is Fairness Only Metric Deep? Evaluating and Addressing Subgroup Gaps in Deep Metric Learning

- Natalie Dullerud, Karsten Roth, Kimia Hamidieh, Nicolas Papernot, Marzyeh Ghassemi
- abstract@open-review(Poster): Deep metric learning (DML) enables learning with less supervision through its emphasis on the similarity structure of representations. There has been much work on improving generalization of DML in settings like zero-shot retrieval, but little is known about its implications for fairness. In this paper, we are the first to evaluate state-of-the-art DML methods trained on imbalanced data, and to show the negative impact these representations have on minority subgroup performance when used for downstream tasks. In this work, we first define fairness in DML through an analysis of three properties of the representation space -- inter-class alignment, intra-class alignment, and uniformity -- and propose \textit{\textbf{finDML}}, the \textit{\textbf{f}}airness \textit{\textbf{i}}n \textit{\textbf{n}}on-balanced \textit{\textbf{DML}} benchmark to characterize representation fairness. Utilizing \textit{finDML}, we find bias in DML representations to propagate to common downstream classification tasks. Surprisingly, this bias is propagated even when training data in the downstream task is re-balanced. To address this problem, we present Partial Attribute De-correlation (\textit{\textbf{\pad}}) to disentangle feature representations from sensitive attributes and reduce performance gaps between subgroups in both embedding space and downstream metrics.

NodePiece: Compositional and Parameter-Efficient Representations of Large Knowledge Graphs

- Mikhail Galkin, Etienne Denis, Jiapeng Wu, William L. Hamilton
- abstract@open-review(Poster): Conventional representation learning algorithms for knowledge graphs (KG) map each entity to a unique embedding vector. Such a shallow lookup results in a linear growth of memory consumption for storing the embedding matrix and incurs high computational costs of working with real-world KGs. Drawing parallels with subword tokenization commonly used in NLP, we explore the landscape of more parameter-efficient node embedding strategies with possibly sublinear memory requirements. To this end, we propose NodePiece, an anchor-based approach to learn a fixed-size entity vocabulary. In NodePiece, a vocabulary of subword/sub-entity units is constructed from anchor nodes in a graph with known relation types. Given such a fixed-size vocabulary, it is possible to bootstrap an encoding and embedding for any entity, including those unseen during training. Experiments show that NodePiece performs competitively in node classification, link prediction, and relation prediction tasks retaining less than 10% of explicit nodes in a graph as anchors and often having 10x fewer parameters. To this end, we show that a NodePiece-enabled model outperforms existing shallow models on a large OGB WikiKG 2 graph having 70x fewer parameters.

Pix2seq: A Language Modeling Framework for Object Detection

- Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, Geoffrey Hinton

- abstract@open-review(Poster): We present Pix2Seq, a simple and generic framework for object detection. Unlike existing approaches that explicitly integrate prior knowledge about the task, we cast object detection as a language modeling task conditioned on the observed pixel inputs. Object descriptions (e.g., bounding boxes and class labels) are expressed as sequences of discrete tokens, and we train a neural network to perceive the image and generate the desired sequence. Our approach is based mainly on the intuition that if a neural network knows about where and what the objects are, we just need to teach it how to read them out. Beyond the use of task-specific data augmentations, our approach makes minimal assumptions about the task, yet it achieves competitive results on the challenging COCO dataset, compared to highly specialized and well optimized detection algorithms.

Particle Stochastic Dual Coordinate Ascent: Exponential convergent algorithm for mean field neural network optimization

- Kazusato Oko, Taiji Suzuki, Atsushi Nitanda, Denny Wu
- abstract@open-review(Poster): We introduce Particle-SDCA, a gradient-based optimization algorithm for two-layer neural networks in the mean field regime that achieves exponential convergence rate in regularized empirical risk minimization. The proposed algorithm can be regarded as an infinite dimensional extension of Stochastic Dual Coordinate Ascent (SDCA) in the probability space: we exploit the convexity of the dual problem, for which the coordinate-wise proximal gradient method can be applied. Our proposed method inherits advantages of the original SDCA, including (i) exponential convergence (with respect to the outer iteration steps), and (ii) better dependency on the sample size and condition number than the full-batch gradient method. One technical challenge in implementing the SDCA update is the intractable integral over the entire parameter space at every step. To overcome this limitation, we propose a tractable \textit{particle method} that approximately solves the dual problem, and an importance re-weighted technique to reduce the computational cost. The convergence rate of our method is verified by numerical experiments.

The Effects of Invertibility on the Representational Complexity of Encoders in Variational Autoencoders

- Divyansh Pareek, Andrej Risteski
- abstract@open-review(Poster): Training and using modern neural-network based latent-variable generative models (like Variational Autoencoders) often require simultaneously training a generative direction along with an inferential (encoding) direction, which approximates the posterior distribution over the latent variables. Thus, the question arises: how complex does the inferential model need to be, in order to be able to accurately model the posterior distribution of a given generative model? In this paper, we identify an important property of the generative map impacting the required size of the encoder. We show that if the generative map is ``strongly invertible" (in a sense we suitably formalize), the inferential model need not be much more complex. Conversely, we prove that there exist non-invertible generative maps, for which the encoding direction needs to be exponentially larger (under standard assumptions in computational complexity). Importantly, we do not require the generative model to be layerwise invertible, which a lot of the related literature assumes and isn't satisfied by many architectures used in practice (e.g. convolution and pooling based networks). Thus, we provide theoretical support for the empirical wisdom that learning deep generative models is harder when data lies on a low-dimensional manifold.

Tracking the risk of a deployed model and detecting harmful distribution shifts

- Aleksandr Podkopaev, Aaditya Ramdas
- abstract@open-review(Poster): When deployed in the real world, machine learning models inevitably encounter changes in the data distribution, and certain---but not all---distribution shifts could result in significant performance degradation. In practice, it may make sense to ignore benign shifts, under which the performance of a deployed model does not degrade substantially, making interventions by a human expert (or model retraining) unnecessary. While several works have developed tests for distribution shifts, these typically either use non-sequential methods, or detect arbitrary shifts (benign or harmful), or both. We argue that a sensible method for firing off a warning has to both (a) detect harmful shifts while ignoring benign ones, and (b) allow continuous monitoring of model performance without increasing the false alarm rate. In this work, we design simple sequential tools for testing if the difference between source (training) and target (test) distributions leads to a significant increase in a risk function of interest, like accuracy or calibration. Recent advances in constructing time-uniform confidence sequences allow efficient aggregation of statistical evidence accumulated during the tracking process. The designed framework is applicable in settings where (some) true labels are revealed after the prediction is performed, or when batches of labels become available in a delayed fashion. We demonstrate the efficacy of the proposed framework through an extensive empirical study on a collection of simulated and real datasets.

Towards Understanding the Robustness Against Evasion Attack on Categorical Data

- Hongyan Bao, Yufei Han, Yujun Zhou, Yun Shen, Xiangliang Zhang
- abstract@open-review(Poster): Characterizing and assessing the adversarial vulnerability of classification models with categorical input has been a practically important, while rarely explored research problem. Our work echoes the challenge by first unveiling the impact factors of adversarial vulnerability of classification models with categorical data based on an information-theoretic adversarial risk analysis about the targeted classifier. Though certifying the robustness of such classification models is intrinsically an NP-hard combinatorial problem, our study shows that the robustness certification can be solved via an efficient greedy exploration of the discrete attack space for any measurable classifiers with a mild smoothness constraint. Our proposed robustness certification framework is instantiated with deep neural network models applied on real-world safety-critic data sources. Our empirical observations confirm the impact of the key adversarial risk factors with categorical input.

Learning Curves for SGD on Structured Features

- Blake Bordelon, Cengiz Pehlevan
- abstract@open-review(Poster): The generalization performance of a machine learning algorithm such as a neural network depends in a non-trivial way on the structure of the data distribution. To analyze the influence of data structure on test loss dynamics, we study an exactly solveable model of stochastic gradient descent (SGD) on the square loss which predicts test error when training on features with arbitrary covariance structure. We solve the theory exactly for both Gaussian features and arbitrary features and we show that the simpler Gaussian model accurately predicts test loss of nonlinear random-feature models and neural networks in the kernel regime trained with SGD on real datasets such as MNIST and CIFAR-10. We show that the optimal batch size at a fixed compute budget is typically small and depends on the feature correlation structure, demonstrating the computational benefits of SGD with small batch sizes. Lastly, we extend our theory to the more usual setting of stochastic gradient descent on a fixed subsampled training set, showing that both training and test error can be accurately predicted in our framework on real data.

NASViT: Neural Architecture Search for Efficient Vision Transformers with Gradient Conflict aware Supernet Training

- Chengyue Gong, Dilin Wang, Meng Li, Xinlei Chen, Zhicheng Yan, Yuandong Tian, qiang liu, Vikas Chandra
- abstract@open-review(Poster): Designing accurate and efficient vision transformers (ViTs) is a highly important but challenging task. Supernet-based one-shot neural architecture search (NAS) enables fast architecture optimization and has achieved state-of-the-art (SOTA) results on convolutional neural networks (CNNs). However, directly applying the supernet-based NAS to optimize ViTs leads to poor performance - even worse compared to training single ViTs. In this work, we observe that the poor performance is due to a gradient conflict issue: the gradients of different sub-networks conflict with that of the supernet more severely in ViTs than CNNs, which leads to early saturation in training and inferior convergence. To alleviate this issue, we propose a series of techniques, including a gradient projection algorithm, a switchable layer scaling design, and a simplified data augmentation and regularization training recipe. The proposed techniques significantly improve the convergence and the performance of all sub-networks. Our discovered hybrid ViT model family, dubbed NASViT, achieves top-1 accuracy from 78.2% to 81.8% on ImageNet from 200M to 800M FLOPs, and outperforms all the prior art CNNs and ViTs, including AlphaNet and LeViT, etc. When transferred to semantic segmentation tasks, NASViTs also outperform previous backbones on

both Cityscape and ADE20K datasets, achieving 73.2% and 37.9% mIoU with only 5G FLOPs, respectively. Code is available at https://github.com/facebookresearch/NASViT.

Graphon based Clustering and Testing of Networks: Algorithms and Theory

- Mahalakshmi Sabanayagam, Leena Chennuru Vankadara, Debarghya Ghoshdastidar
- abstract@open-review(Poster): Network-valued data are encountered in a wide range of applications, and pose challenges in learning due to their complex structure and absence of vertex correspondence. Typical examples of such problems include classification or grouping of protein structures and social networks. Various methods, ranging from graph kernels to graph neural networks, have been proposed that achieve some success in graph classification problems. However, most methods have limited theoretical justification, and their applicability beyond classification remains unexplored. In this work, we propose methods for clustering multiple graphs, without vertex correspondence, that are inspired by the recent literature on estimating graphons---symmetric functions corresponding to infinite vertex limit of graphs. We propose a novel graph distance based on sorting-and-smoothing graphon estimators. Using the proposed graph distance, we present two clustering algorithms and show that they achieve state-of-the-art results. We prove the statistical consistency of both algorithms under Lipschitz assumptions on the graph degrees. We further study the applicability of the proposed distance for graph two-sample testing problems.

Network Augmentation for Tiny Deep Learning

- Han Cai, Chuang Gan, Ji Lin, song han
- abstract@open-review(Poster): We introduce Network Augmentation (NetAug), a new training method for improving the performance of tiny neural networks. Existing regularization techniques (e.g., data augmentation, dropout) have shown much success on large neural networks by adding noise to overcome over-fitting. However, we found these techniques hurt the performance of tiny neural networks. We argue that training tiny models are different from large models: rather than augmenting the data, we should augment the model, since tiny models tend to suffer from under-fitting rather than over-fitting due to limited capacity. To alleviate this issue, NetAug augments the network (reverse dropout) instead of inserting noise into the dataset or the network. It puts the tiny model into larger models and encourages it to work as a sub-model of larger models to get extra supervision, in addition to functioning as an independent model. At test time, only the tiny model is used for inference, incurring zero inference overhead. We demonstrate the effectiveness of NetAug on image classification and object detection. NetAug consistently improves the performance of tiny models, achieving up to 2.2% accuracy improvement on ImageNet. On object detection, achieving the same level of performance, NetAug requires 41% fewer MACs on Pascal VOC and 38% fewer MACs on COCO than the baseline.

Bridging the Gap: Providing Post-Hoc Symbolic Explanations for Sequential Decision-Making Problems with Inscrutable Representations

- Sarath Sreedharan, Utkarsh Soni, Mudit Verma, Siddharth Srivastava, Subbarao Kambhampati
- abstract@open-review(Poster): As increasingly complex AI systems are introduced into our daily lives, it becomes important for such systems to be capable of explaining the rationale for their decisions and allowing users to contest these decisions. A significant hurdle to allowing for such explanatory dialogue could be the {\em vocabulary mismatch} between the user and the AI system. This paper introduces methods for providing contrastive explanations in terms of user-specified concepts for sequential decision-making settings where the system's model of the task may be best represented as an inscrutable model. We do this by building partial symbolic models of a local approximation of the task that can be leveraged to answer the user queries. We test these methods on a popular Atari game (Montezuma's Revenge) and variants of Sokoban (a well-known planning benchmark) and report the results of user studies to evaluate whether people find explanations generated in this form useful.

Distributional Reinforcement Learning with Monotonic Splines

- Yudong Luo, Guiliang Liu, Haonan Duan, Oliver Schulte, Pascal Poupart
- abstract@open-review(Poster): Distributional Reinforcement Learning (RL) differs from traditional RL by estimating the distribution over returns to capture the intrinsic uncertainty of MDPs. One key challenge in distributional RL lies in how to parameterize the quantile function when minimizing the Wasserstein metric of temporal differences. Existing algorithms use step functions or piecewise linear functions. In this paper, we propose to learn smooth continuous quantile functions represented by monotonic rational-quadratic splines, which also naturally solve the quantile crossing problem. Experiments in stochastic environments show that a dense estimation for quantile functions enhances distributional RL in terms of faster empirical convergence and higher rewards in most cases.

Toward Faithful Case-based Reasoning through Learning Prototypes in a Nearest Neighbor-friendly Space.

- Seyed Omid Davoudi, Majid Komeili
- abstract@open-review(Poster): Recent advances in machine learning have brought opportunities for the ever-increasing use of AI in the real world. This has created concerns about the black-box nature of many of the most recent machine learning approaches. In this work, we propose an interpretable neural network that leverages metric and prototype learning for classification tasks. It encodes its own explanations and provides an improved case-based reasoning through learning prototypes in an embedding space learned by a probabilistic nearest neighbor rule. Through experiments, we demonstrated the effectiveness of the proposed method in both performance and the accuracy of the explanations provided.

Augmented Sliced Wasserstein Distances

- Xiongjie Chen, Yongxin Yang, Yunpeng Li
- abstract@open-review(Poster): While theoretically appealing, the application of the Wasserstein distance to large-scale machine learning problems has been hampered by its prohibitive computational cost. The sliced Wasserstein distance and its variants improve the computational efficiency through the random projection, yet they suffer from low accuracy if the number of projections is not sufficiently large, because the majority of projections result in trivially small values. In this work, we propose a new family of distance metrics, called augmented sliced Wasserstein distances (ASWDs), constructed by first mapping samples to higher-dimensional hypersurfaces parameterized by neural networks. It is derived from a key observation that (random) linear projections of samples residing on these hypersurfaces would translate to much more flexible nonlinear projections in the original sample space, so they can capture complex structures of the data distribution. We show that the hypersurfaces can be optimized by gradient ascent efficiently. We provide the condition under which the ASWD is a valid metric and show that this can be obtained by an injective neural network architecture. Numerical results demonstrate that the ASWD significantly outperforms other Wasserstein variants for both synthetic and real-world problems.

Relational Learning with Variational Bayes

- Kuang-Hung Liu
- abstract@open-review(Poster): In psychology, relational learning refers to the ability to recognize and respond to relationship among objects irrespective of the nature of those objects. Relational learning has long been recognized as a hallmark of human cognition and a key question in artificial intelligence research. In this work, we propose an unsupervised learning method for addressing the relational learning problem where we learn the underlying

relationship between a pair of data irrespective of the nature of those data. The central idea of the proposed method is to encapsulate the relational learning problem with a probabilistic graphical model in which we perform inference to learn about data relationship and other relational processing tasks.

Provably Robust Adversarial Examples

- Dimitar Iliev Dimitrov, Gagandeep Singh, Timon Gehr, Martin Vechev
- abstract@open-review(Poster): We introduce the concept of provably robust adversarial examples for deep neural networks â€“ connected input regions constructed from standard adversarial examples which are guaranteed to be robust to a set of real-world perturbations (such as changes in pixel intensity and geometric transformations). We present a novel method called PARADE for generating these regions in a scalable manner which works by iteratively refining the region initially obtained via sampling until a refined region is certified to be adversarial with existing state-of-the-art verifiers. At each step, a novel optimization procedure is applied to maximize the region's volume under the constraint that the convex relaxation of the network behavior with respect to the region implies a chosen bound on the certification objective. Our experimental evaluation shows the effectiveness of PARADE: it successfully finds large provably robust regions including ones containing $\approx 10^{573}$ adversarial examples for pixel intensity and $\approx 10^{599}$ for geometric perturbations. The provability enables our robust examples to be significantly more effective against state-of-the-art defenses based on randomized smoothing than the individual attacks used to construct the regions.

Joint Shapley values: a measure of joint feature importance

- Chris Harris, Richard Pymar, Colin Rowat
- abstract@open-review(Poster): The Shapley value is one of the most widely used measures of feature importance partly as it measures a feature's average effect on a model's prediction. We introduce joint Shapley values, which directly extend Shapley's axioms and intuitions: joint Shapley values measure a set of features' average effect on a model's prediction. We prove the uniqueness of joint Shapley values, for any order of explanation. Results for games show that joint Shapley values present different insights from existing interaction indices, which assess the effect of a feature within a set of features. The joint Shapley values seem to provide sensible results in ML attribution problems. With binary features, we present a presence-adjusted global value that is more consistent with local intuitions than the usual approach.

Low-Budget Active Learning via Wasserstein Distance: An Integer Programming Approach

- Rafid Mahmood, Sanja Fidler, Marc T Law
- abstract@open-review(Poster): Active learning is the process of training a model with limited labeled data by selecting a core subset of an unlabeled data pool to label. The large scale of data sets used in deep learning forces most sample selection strategies to employ efficient heuristics. This paper introduces an integer optimization problem for selecting a core set that minimizes the discrete Wasserstein distance from the unlabeled pool. We demonstrate that this problem can be tractably solved with a Generalized Benders Decomposition algorithm. Our strategy uses high-quality latent features that can be obtained by unsupervised learning on the unlabeled pool. Numerical results on several data sets show that our optimization approach is competitive with baselines and particularly outperforms them in the low budget regime where less than one percent of the data set is labeled.

Efficient Self-supervised Vision Transformers for Representation Learning

- Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, Jianfeng Gao
- abstract@open-review(Poster): This paper investigates two techniques for developing efficient self-supervised vision transformers (EsViT) for visual representation learning. First, we show through a comprehensive empirical study that multi-stage architectures with sparse self-attentions can significantly reduce modeling complexity but with a cost of losing the ability to capture fine-grained correspondences between image regions. Second, we propose a new pre-training task, non-contrastive region-matching, which allows the model to capture fine-grained region dependencies and as a result significantly improves the quality of the learned vision representations. Our results show that combining the two techniques, EsViT achieves 81.3% top-1 on the ImageNet linear probe evaluation, outperforming prior arts with around an order magnitude of higher throughput. When transferring to downstream linear classification tasks, EsViT outperforms its supervised counterpart on 17 out of 18 datasets. The code and pre-trained models are released at: https://github.com/microsoft/esvit

Visual Representation Learning Does Not Generalize Strongly Within the Same Domain

- Lukas Schott, Julius Von KÃ¼gelgen, Frederik TrÃ¤uble, Peter Vincent Gehler, Chris Russell, Matthias Bethge, Bernhard SchÃ¶lkopf, Francesco Locatello, Wieland Brendel
- abstract@open-review(Poster): An important component for generalization in machine learning is to uncover underlying latent factors of variation as well as the mechanism through which each factor acts in the world. In this paper, we test whether 17 unsupervised, weakly supervised, and fully supervised representation learning approaches correctly infer the generative factors of variation in simple datasets (dSprites, Shapes3D, MPI3D) from controlled environments, and on our contributed CelebGlow dataset. In contrast to prior robustness work that introduces novel factors of variation during test time, such as blur or other (un)structured noise, we here recompose, interpolate, or extrapolate only existing factors of variation from the training data set (e.g., small and medium-sized objects during training and large objects during testing). Models that learn the correct mechanism should be able to generalize to this benchmark. In total, we train and test 2000+ models and observe that all of them struggle to learn the underlying mechanism regardless of supervision signal and architectural bias. Moreover, the generalization capabilities of all tested models drop significantly as we move from artificial datasets towards more realistic real-world datasets. Despite their inability to identify the correct mechanism, the models are quite modular as their ability to infer other in-distribution factors remains fairly stable, providing only a single factor is out-of-distribution. These results point to an important yet understudied problem of learning mechanistic models of observations that can facilitate generalization.

Hidden Convexity of Wasserstein GANs: Interpretable Generative Models with Closed-Form Solutions

- Arda Sahiner, Tolga Ergen, Batu Ozturkler, Burak Bartan, John M. Pauly, Morteza Mardani, Mert Pilanci
- abstract@open-review(Poster): Generative Adversarial Networks (GANs) are commonly used for modeling complex distributions of data. Both the generators and discriminators of GANs are often modeled by neural networks, posing a non-transparent optimization problem which is non-convex and non-concave over the generator and discriminator, respectively. Such networks are often heuristically optimized with gradient descent-ascent (GDA), but it is unclear whether the optimization problem contains any saddle points, or whether heuristic methods can find them in practice. In this work, we analyze the training of Wasserstein GANs with two-layer neural network discriminators through the lens of convex duality, and for a variety of generators expose the conditions under which Wasserstein GANs can be solved exactly with convex optimization approaches, or can be represented as convex-concave games. Using this convex duality interpretation, we further demonstrate the impact of different activation functions of the discriminator. Our observations are verified with numerical results demonstrating the power of the convex interpretation, with an application in progressive training of convex architectures corresponding to linear generators and quadratic-activation discriminators for CelebA image generation. The code for our experiments is available at https://github.com/ardasahiner/ProCoGAN.

Memory Augmented Optimizers for Deep Learning

- Paul-Aymeric Martin McRae, Prasanna Parthasarathi, Mido Assran, Sarath Chandar

- abstract@open-review(Poster): Popular approaches for minimizing loss in data-driven learning often involve an abstraction or an explicit retention of the history of gradients for efficient parameter updates. The aggregated history of gradients nudges the parameter updates in the right direction even when the gradients at any given step are not informative. Although the history of gradients summarized in meta-parameters or explicitly stored in memory has been shown effective in theory and practice, the question of whether all or only a subset of the gradients in the history are sufficient in deciding the parameter updates remains unanswered. In this paper, we propose a framework of memory-augmented gradient descent optimizers that retain a limited view of their gradient history in their internal memory. Such optimizers scale well to large real-life datasets, and our experiments show that the memory augmented extensions of standard optimizers enjoy accelerated convergence and improved performance on a majority of computer vision and language tasks that we considered. Additionally, we prove that the proposed class of optimizers with fixed-size memory converge under assumptions of strong convexity, regardless of which gradients are selected or how they are linearly combined to form the update step.

Orchestrated Value Mapping for Reinforcement Learning

- Mehdi Fatemi, Arash Tavakoli
- abstract@open-review(Poster): We present a general convergent class of reinforcement learning algorithms that is founded on two distinct principles: (1) mapping value estimates to a different space using arbitrary functions from a broad class, and (2) linearly decomposing the reward signal into multiple channels. The first principle enables incorporating specific properties into the value estimator that can enhance learning. The second principle, on the other hand, allows for the value function to be represented as a composition of multiple utility functions. This can be leveraged for various purposes, e.g. dealing with highly varying reward scales, incorporating a priori knowledge about the sources of reward, and ensemble learning. Combining the two principles yields a general blueprint for instantiating convergent algorithms by orchestrating diverse mapping functions over multiple reward channels. This blueprint generalizes and subsumes algorithms such as Q-Learning, Log Q-Learning, and Q-Decomposition. In addition, our convergence proof for this general class relaxes certain required assumptions in some of these algorithms. Based on our theory, we discuss several interesting configurations as special cases. Finally, to illustrate the potential of the design space that our theory opens up, we instantiate a particular algorithm and evaluate its performance on the Atari suite.

Learning to Generalize across Domains on Single Test Samples

- Zehao Xiao, Xiantong Zhen, Ling Shao, Cees G. M. Snoek
- abstract@open-review(Poster): We strive to learn a model from a set of source domains that generalizes well to unseen target domains. The main challenge in such a domain generalization scenario is the unavailability of any target domain data during training, resulting in the learned model not being explicitly adapted to the unseen target domains. We propose learning to generalize across domains on single test samples. We leverage a meta-learning paradigm to learn our model to acquire the ability of adaptation with single samples at training time so as to further adapt itself to each single test sample at test time. We formulate the adaptation to the single test sample as a variational Bayesian inference problem, which incorporates the test sample as a conditional into the generation of model parameters. The adaptation to each test sample requires only one feed-forward computation at test time without any fine-tuning or self-supervised training on additional data from the unseen domains. Extensive ablation studies demonstrate that our model learns the ability to adapt models to each single sample by mimicking domain shifts during training. Further, our model achieves at least comparable -- and often better -- performance than state-of-the-art methods on multiple benchmarks for domain generalization.

Prototype memory and attention mechanisms for few shot image generation

- Tianqin Li, Zijie Li, Andrew Luo, Harold Rockwell, Amir Barati Farimani, Tai Sing Lee
- abstract@open-review(Poster): Recent discoveries indicate that the neural codes in the primary visual cortex (V1) of macaque monkeys are complex, diverse and sparse. This leads us to ponder the computational advantages and functional role of these â€œgrandmother cells." Here, we propose that such cells can serve as prototype memory priors that bias and shape the distributed feature processing within the image generation process in the brain. These memory prototypes are learned by momentum online clustering and are utilized via a memory-based attention operation, which we define as Memory Concept Attention (MoCA). To test our proposal, we show in a few-shot image generation task, that having a prototype memory during attention can improve image synthesis quality, learn interpretable visual concept clusters, as well as improve the robustness of the model. Interestingly, we also find that our attentional memory mechanism can implicitly modify the horizontal connections by updating the transformation into the prototype embedding space for self-attention. Insofar as GANs can be seen as plausible models for reasoning about the top-down synthesis in the analysis-by-synthesis loop of the hierarchical visual cortex, our findings demonstrate a plausible computational role for these â€œprototype concept" neurons in visual processing in the brain.

TPU-GAN: Learning temporal coherence from dynamic point cloud sequences

- Zijie Li, Tianqin Li, Amir Barati Farimani
- abstract@open-review(Poster): Point cloud sequence is an important data representation that provides flexible shape and motion information. Prior work demonstrates that incorporating scene flow information into loss can make model learn temporally coherent feature spaces. However, it is prohibitively expensive to acquire point correspondence information across frames in real-world environments. In this work, we propose a super-resolution generative adversarial network (GAN) for upsampling dynamic point cloud sequences, which does not require point correspondence annotation. Our model, Temporal Point cloud Upsampling GAN (TPU-GAN), can implicitly learn the underlying temporal coherence from point cloud sequence, which in turn guides the generator to produce temporally coherent output. In addition, we propose a learnable masking module to adapt upsampling ratio according to the point distribution. We conduct extensive experiments on point cloud sequences from two different domains: particles in the fluid dynamical system and human action scanned data. The quantitative and qualitative evaluation demonstrates the effectiveness of our method on upsampling tasks as well as learning temporal coherence from irregular point cloud sequences.

A First-Occupancy Representation for Reinforcement Learning

- Ted Moskovitz, Spencer R Wilson, Maneesh Sahani
- abstract@open-review(Poster): Both animals and artificial agents benefit from state representations that support rapid transfer of learning across tasks and which enable them to efficiently traverse their environments to reach rewarding states. The successor representation (SR), which measures the expected cumulative, discounted state occupancy under a fixed policy, enables efficient transfer to different reward structures in an otherwise constant Markovian environment and has been hypothesized to underlie aspects of biological behavior and neural activity. However, in the real world, rewards may only be available for consumption once, may shift location, or agents may simply aim to reach goal states as rapidly as possible without the constraint of artificially imposed task horizons. In such cases, the most behaviorally-relevant representation would carry information about when the agent was likely to first reach states of interest, rather than how often it should expect to visit them over a potentially infinite time span. To reflect such demands, we introduce the first-occupancy representation (FR), which measures the expected temporal discount to the first time a state is accessed. We demonstrate that the FR facilitates exploration, the selection of efficient paths to desired states, allows the agent, under certain conditions, to plan provably optimal trajectories defined by a sequence of subgoals, and induces similar behavior to animals avoiding threatening stimuli.

Deep ReLU Networks Preserve Expected Length

- Boris Hanin, Ryan S Jeong, David Rolnick

- abstract@open-review(Poster): Assessing the complexity of functions computed by a neural network helps us understand how the network will learn and generalize. One natural measure of complexity is how the network distorts length - if the network takes a unit-length curve as input, what is the length of the resulting curve of outputs? It has been widely believed that this length grows exponentially in network depth. We prove that in fact this is not the case: the expected length distortion does not grow with depth, and indeed shrinks slightly, for ReLU networks with standard random initialization. We also generalize this result by proving upper bounds both for higher moments of the length distortion and for the distortion of higher-dimensional volumes. These theoretical results are corroborated by our experiments.

Phenomenology of Double Descent in Finite-Width Neural Networks

- Sidak Pal Singh, Aurelien Lucchi, Thomas Hofmann, Bernhard SchÃ¶lkopf
- abstract@open-review(Poster): `Double descent' delineates the generalization behaviour of models depending on the regime they belong to: under- or over-parameterized. The current theoretical understanding behind the occurrence of this phenomenon is primarily based on linear and kernel regression models --- with informal parallels to neural networks via the Neural Tangent Kernel. Therefore such analyses do not adequately capture the mechanisms behind double descent in finite-width neural networks, as well as, disregard crucial components --- such as the choice of the loss function. We address these shortcomings by leveraging influence functions in order to derive suitable expressions of the population loss and its lower bound, while imposing minimal assumptions on the form of the parametric model. Our derived bounds bear an intimate connection with the spectrum of the Hessian at the optimum, and importantly, exhibit a double descent behaviour at the interpolation threshold. Building on our analysis, we further investigate how the loss function affects double descent --- and thus uncover interesting properties of neural networks and their Hessian spectra near the interpolation threshold.

How Attentive are Graph Attention Networks?

- Shaked Brody, Uri Alon, Eran Yahav
- abstract@open-review(Poster): Graph Attention Networks (GATs) are one of the most popular GNN architectures and are considered as the state-of-the-art architecture for representation learning with graphs. In GAT, every node attends to its neighbors given its own representation as the query. However, in this paper we show that GAT computes a very limited kind of attention: the ranking of the attention scores is unconditioned on the query node. We formally define this restricted kind of attention as static attention and distinguish it from a strictly more expressive dynamic attention. Because GATs use a static attention mechanism, there are simple graph problems that GAT cannot express: in a controlled problem, we show that static attention hinders GAT from even fitting the training data. To remove this limitation, we introduce a simple fix by modifying the order of operations and propose GATv2: a dynamic graph attention variant that is strictly more expressive than GAT. We perform an extensive evaluation and show that GATv2 outperforms GAT across 12 OGB and other benchmarks while we match their parametric costs. Our code is available at https://github.com/tech-srl/how_attentive_are_gats . GATv2 is available as part of the PyTorch Geometric library, the Deep Graph Library, and the TensorFlow GNN library.

Learning Transferable Reward for Query Object Localization with Policy Adaptation

- Tingfeng Li, Shaobo Han, Martin Renqiang Min, Dimitris N. Metaxas
- abstract@open-review(Poster): We propose a reinforcement learning based approach to query object localization, for which an agent is trained to localize objects of interest specified by a small exemplary set. We learn a transferable reward signal formulated using the exemplary set by ordinal metric learning. Our proposed method enables test-time policy adaptation to new environments where the reward signals are not readily available, and outperforms fine-tuning approaches that are limited to annotated images. In addition, the transferable reward allows repurposing the trained agent from one specific class to another class. Experiments on corrupted MNIST, CU-Birds, and COCO datasets demonstrate the effectiveness of our approach.

CKConv: Continuous Kernel Convolution For Sequential Data

- David W. Romero, Anna Kuzina, Erik J Bekkers, Jakub Mikolaj Tomczak, Mark Hoogendoorn
- abstract@open-review(Poster): Conventional neural architectures for sequential data present important limitations. Recurrent neural networks suffer from exploding and vanishing gradients, small effective memory horizons, and must be trained sequentially. Convolutional neural networks cannot handle sequences of unknown size and their memory horizon must be defined a priori. In this work, we show that these problems can be solved by formulating the convolutional kernels of CNNs as continuous functions. The resulting Continuous Kernel Convolution (CKConv) handles arbitrarily long sequences in a parallel manner, within a single operation, and without relying on any form of recurrence. We show that Continuous Kernel Convolutional Networks (CKCNNs) obtain state-of-the-art results in multiple datasets, e.g., permuted MNIST, and, thanks to their continuous nature, are able to handle non-uniformly sampled datasets and irregularly-sampled data natively. CKCNNs match or perform better than neural ODEs designed for these purposes in a faster and simpler manner.

Towards Empirical Sandwich Bounds on the Rate-Distortion Function

- Yibo Yang, Stephan Mandt
- abstract@open-review(Poster): Rate-distortion (R-D) function, a key quantity in information theory, characterizes the fundamental limit of how much a data source can be compressed subject to a fidelity criterion, by any compression algorithm. As researchers push for ever-improving compression performance, establishing the R-D function of a given data source is not only of scientific interest, but also reveals the possible room for improvement in existing compression algorithms. Previous work on this problem relied on distributional assumptions on the data source (Gibson, 2017) or only applied to discrete data (Blahut, 1972; Arimoto, 1972). By contrast, this paper makes the first attempt at an algorithm for sandwiching the R-D function of a general (not necessarily discrete) source requiring only i.i.d. data samples. We estimate R-D sandwich bounds for a variety of artificial and real-world data sources, in settings far beyond the feasibility of any known method, and shed light on the optimality of neural data compression (BallÃ© et al., 2021; Yang et al., 2022). Our R-D upper bound on natural images indicates theoretical room for improving state-of-the-art image compression methods by at least one dB in PSNR at various bitrates. Our data and code can be found at https://github.com/mandt-lab/RD-sandwich.

Pareto Policy Adaptation

- Panagiotis Kyriakis, Jyotirmoy Deshmukh, Paul Bogdan
- abstract@open-review(Poster): We present a policy gradient method for Multi-Objective Reinforcement Learning under unknown, linear preferences. By enforcing Pareto stationarity, a first-order condition for Pareto optimality, we are able to design a simple policy gradient algorithm that approximates the Pareto front and infers the unknown preferences. Our method relies on a projected gradient descent solver that identifies common ascent directions for all objectives. Leveraging the solution of that solver, we introduce Pareto Policy Adaptation (PPA), a loss function that adapts the policy to be optimal with respect to any distribution over preferences. PPA uses implicit differentiation to back-propagate the loss gradient bypassing the operations of the projected gradient descent solver. Our approach is straightforward, easy to implement and can be used with all existing policy gradient and actor-critic methods. We evaluate our method in a series of reinforcement learning tasks

Fair Normalizing Flows

- Mislav Balunovic, Anian Ruoss, Martin Vechev

- abstract@open-review(Poster): Fair representation learning is an attractive approach that promises fairness of downstream predictors by encoding sensitive data. Unfortunately, recent work has shown that strong adversarial predictors can still exhibit unfairness by recovering sensitive attributes from these representations. In this work, we present Fair Normalizing Flows (FNF), a new approach offering more rigorous fairness guarantees for learned representations. Specifically, we consider a practical setting where we can estimate the probability density for sensitive groups. The key idea is to model the encoder as a normalizing flow trained to minimize the statistical distance between the latent representations of different groups. The main advantage of FNF is that its exact likelihood computation allows us to obtain guarantees on the maximum unfairness of any potentially adversarial downstream predictor. We experimentally demonstrate the effectiveness of FNF in enforcing various group fairness notions, as well as other attractive properties such as interpretability and transfer learning, on a variety of challenging real-world datasets.

The Convex Geometry of Backpropagation: Neural Network Gradient Flows Converge to Extreme Points of the Dual Convex Program

- Yifei Wang, Mert Pilanci
- abstract@open-review(Poster): We study non-convex subgradient flows for training two-layer ReLU neural networks from a convex geometry and duality perspective. We characterize the implicit bias of unregularized non-convex gradient flow as convex regularization of an equivalent convex model. We then show that the limit points of non-convex subgradient flows can be identified via primal-dual correspondence in this convex optimization problem. Moreover, we derive a sufficient condition on the dual variables which ensures that the stationary points of the non-convex objective are the KKT points of the convex objective, thus proving convergence of non-convex gradient flows to the global optimum. For a class of regular training data distributions such as orthogonal separable data, we show that this sufficient condition holds. Therefore, non-convex gradient flows in fact converge to optimal solutions of a convex optimization problem. We present numerical results verifying the predictions of our theory for non-convex subgradient descent.

Adaptive Wavelet Transformer Network for 3D Shape Representation Learning

- Hao Huang, Yi Fang
- abstract@open-review(Poster): We present a novel method for 3D shape representation learning using multi-scale wavelet decomposition. Previous works often decompose 3D shapes into complementary components in spatial domain at a single scale. In this work, we study to decompose 3D shapes into sub-bands components in frequency domain at multiple scales, resulting in a hierarchical decomposition tree in a principled manner rooted in multi-resolution wavelet analysis. Specifically, we propose Adaptive Wavelet Transformer Network (AWT-Net) that firstly generates approximation or detail wavelet coefficients per point, classifying each point into high or low sub-bands components, using lifting scheme at multiple scales recursively and hierarchically. Then, AWT-Net exploits Transformer to enhance the original shape features by querying and fusing features from different but integrated sub-bands. The wavelet coefficients can be learned without direct supervision on coefficients, and AWT-Net is fully differentiable and can be learned in an end-to-end fashion. Extensive experiments demonstrate that AWT-Net achieves competitive performance on 3D shape classification and segmentation benchmarks.

On the Convergence of mSGD and AdaGrad for Stochastic Optimization

- ruinan Jin, Yu Xing, Xingkang He
- abstract@open-review(Poster): As one of the most fundamental stochastic optimization algorithms, stochastic gradient descent (SGD) has been intensively developed and extensively applied in machine learning in the past decade. There have been some modified SGD-type algorithms, which outperform the SGD in many competitions and applications in terms of convergence rate and accuracy, such as momentum-based SGD (mSGD) and adaptive gradient algorithm (AdaGrad). Despite these empirical successes, the theoretical properties of these algorithms have not been well established due to technical difficulties. With this motivation, we focus on convergence analysis of mSGD and AdaGrad for any smooth (possibly non-convex) loss functions in stochastic optimization. First, we prove that the iterates of mSGD are asymptotically convergent to a connected set of stationary points with probability one, which is more general than existing works on subsequence convergence or convergence of time averages. Moreover, we prove that the loss function of mSGD decays at a certain rate faster than that of SGD. In addition, we prove the iterates of AdaGrad are asymptotically convergent to a connected set of stationary points with probability one. Also, this result extends the results from the literature on subsequence convergence and the convergence of time averages. Despite the generality of the above convergence results, we have relaxed some assumptions of gradient noises, convexity of loss functions, as well as boundedness of iterates.

Likelihood Training of SchrÃ¶dinger Bridge using Forward-Backward SDEs Theory

- Tianrong Chen, Guan-Horng Liu, Evangelos Theodorou
- abstract@open-review(Poster): SchrÃ¶dinger Bridge (SB) is an entropy-regularized optimal transport problem that has received increasing attention in deep generative modeling for its mathematical flexibility compared to the Scored-based Generative Model (SGM). However, it remains unclear whether the optimization principle of SB relates to the modern training of deep generative models, which often rely on constructing log-likelihood objectives.This raises questions on the suitability of SB models as a principled alternative for generative applications. In this work, we present a novel computational framework for likelihood training of SB models grounded on Forward-Backward Stochastic Differential Equations Theory â€“ a mathematical methodology appeared in stochastic optimal control that transforms the optimality condition of SB into a set of SDEs. Crucially, these SDEs can be used to construct the likelihood objectives for SB that, surprisingly, generalizes the ones for SGM as special cases. This leads to a new optimization principle that inherits the same SB optimality yet without losing applications of modern generative training techniques, and we show that the resulting training algorithm achieves comparable results on generating realistic images on MNIST, CelebA, and CIFAR10. Our code is available at https://github.com/ghliu/SB-FBSDE.

Imitation Learning from Observations under Transition Model Disparity

- Tanmay Gangwani, Yuan Zhou, Jian Peng
- abstract@open-review(Poster): Learning to perform tasks by leveraging a dataset of expert observations, also known as imitation learning from observations (ILO), is an important paradigm for learning skills without access to the expert reward function or the expert actions. We consider ILO in the setting where the expert and the learner agents operate in different environments, with the source of the discrepancy being the transition dynamics model. Recent methods for scalable ILO utilize adversarial learning to match the state-transition distributions of the expert and the learner, an approach that becomes challenging when the dynamics are dissimilar. In this work, we propose an algorithm that trains an intermediary policy in the learner environment and uses it as a surrogate expert for the learner. The intermediary policy is learned such that the state transitions generated by it are close to the state transitions in the expert dataset. To derive a practical and scalable algorithm, we employ concepts from prior work on estimating the support of a probability distribution. Experiments using MuJoCo locomotion tasks highlight that our method compares favorably to the baselines for ILO with transition dynamics mismatch.

MCMC Should Mix: Learning Energy-Based Model with Neural Transport Latent Space MCMC

- Erik Nijkamp, Ruiqi Gao, Pavel Sountsov, Srinivas Vasudevan, Bo Pang, Song-Chun Zhu, Ying Nian Wu
- abstract@open-review(Poster): Learning energy-based model (EBM) requires MCMC sampling of the learned model as an inner loop of the learning algorithm. However, MCMC sampling of EBMs in high-dimensional data space is generally not mixing, because the energy function, which is usually parametrized by deep network, is highly multi-modal in the data space. This is a serious handicap for both theory and practice of EBMs. In this paper, we propose to learn EBM with a flow-based model (or in general latent variable model) serving as a backbone, so that the EBM is a correction or an

exponential tilting of the flow-based model. We show that the model has a particularly simple form in the space of the latent variables of the generative model, and MCMC sampling of the EBM in the latent space mixes well and traverses modes in the data space. This enables proper sampling and learning of EBMs.

Autonomous Learning of Object-Centric Abstractions for High-Level Planning

- Steven James, Benjamin Rosman, George Konidaris
- abstract@open-review(Poster): We propose a method for autonomously learning an object-centric representation of a continuous and high-dimensional environment that is suitable for planning. Such representations can immediately be transferred between tasks that share the same types of objects, resulting in agents that require fewer samples to learn a model of a new task. We first demonstrate our approach on a 2D crafting domain consisting of numerous objects where the agent learns a compact, lifted representation that generalises across objects. We then apply it to a series of Minecraft tasks to learn object-centric representations and object types - directly from pixel data - that can be leveraged to solve new tasks quickly. The resulting learned representations enable the use of a task-level planner, resulting in an agent capable of transferring learned representations to form complex, long-term plans.

A fast and accurate splitting method for optimal transport: analysis and implementation

- Vien V. Mai, Jacob LindbÃ¤ck, Mikael Johansson
- abstract@open-review(Poster): We develop a fast and reliable method for solving large-scale optimal transport (OT) problems at an unprecedented combination of speed and accuracy. Built on the celebrated Douglas-Rachford splitting technique, our method tackles the original OT problem directly instead of solving an approximate regularized problem, as many state-of-the-art techniques do. This allows us to provide sparse transport plans and avoid numerical issues of methods that use entropic regularization. The algorithm has the same cost per iteration as the popular Sinkhorn method, and each iteration can be executed efficiently, in parallel. The proposed method enjoys an iteration complexity $O(1/\epsilon)$ compared to the best-known $O(1/\epsilon^2)$ of the Sinkhorn method. In addition, we establish a linear convergence rate for our formulation of the OT problem. We detail an efficient GPU implementation of the proposed method that maintains a primal-dual stopping criterion at no extra cost. Substantial experiments demonstrate the effectiveness of our method, both in terms of computation times and robustness.

Implicit Bias of MSE Gradient Optimization in Underparameterized Neural Networks

- Benjamin Bowman, Guido Montufar
- abstract@open-review(Poster): We study the dynamics of a neural network in function space when optimizing the mean squared error via gradient flow. We show that in the underparameterized regime the network learns eigenfunctions of an integral operator T_K determined by the Neural Tangent Kernel at rates corresponding to their eigenvalues. For example, for uniformly distributed data on the sphere $S^{d - 1}$ and rotation invariant weight distributions, the eigenfunctions of T_K are the spherical harmonics. Our results can be understood as describing a spectral bias in the underparameterized regime. The proofs use the concept of ``Damped Deviations" where deviations of the NTK matter less for eigendirections with large eigenvalues. Aside from the underparameterized regime, the damped deviations point-of-view allows us to extend certain results in the literature in the overparameterized setting.

Discovering Latent Concepts Learned in BERT

- Fahim Dalvi, Abdul Rafae Khan, Firoj Alam, Nadir Durrani, Jia Xu, Hassan Sajjad
- abstract@open-review(Poster): A large number of studies that analyze deep neural network models and their ability to encode various linguistic and non-linguistic concepts provide an interpretation of the inner mechanics of these models. The scope of the analyses is limited to pre-defined concepts that reinforce the traditional linguistic knowledge and do not reflect on how novel concepts are learned by the model. We address this limitation by discovering and analyzing latent concepts learned in neural network models in an unsupervised fashion and provide interpretations from the model's perspective. In this work, we study: i) what latent concepts exist in the pre-trained BERT model, ii) how the discovered latent concepts align or diverge from classical linguistic hierarchy and iii) how the latent concepts evolve across layers. Our findings show: i) a model learns novel concepts (e.g. animal categories and demographic groups), which do not strictly adhere to any pre-defined categorization (e.g. POS, semantic tags), ii) several latent concepts are based on multiple properties which may include semantics, syntax, and morphology, iii) the lower layers in the model dominate in learning shallow lexical concepts while the higher layers learn semantic relations and iv) the discovered latent concepts highlight potential biases learned in the model. We also release a novel BERT ConceptNet dataset consisting of 174 concept labels and 1M annotated instances.

The Role of Permutation Invariance in Linear Mode Connectivity of Neural Networks

- Rahim Entezari, Hanie Sedghi, Olga Saukh, Behnam Neyshabur
- abstract@open-review(Poster): In this paper, we conjecture that if the permutation invariance of neural networks is taken into account, SGD solutions will likely have no barrier in the linear interpolation between them. Although it is a bold conjecture, we show how extensive empirical attempts fall short of refuting it. We further provide a preliminary theoretical result to support our conjecture. Our conjecture has implications for the lottery ticket hypothesis, distributed training, and ensemble methods. The source code is available at \url{https://github.com/rahimentezari/PermutationInvariance}.

Data Poisoning Wonâ€™t Save You From Facial Recognition

- Evani Radiya-Dixit, Sanghyun Hong, Nicholas Carlini, Florian Tramer
- abstract@open-review(Poster): Data poisoning has been proposed as a compelling defense against facial recognition models trained on Web-scraped pictures. Users can perturb images they post online, so that models will misclassify future (unperturbed) pictures.

We demonstrate that this strategy provides a false sense of security, as it ignores an inherent asymmetry between the parties: users' pictures are perturbed once and for all before being published (at which point they are scraped) and must thereafter fool all future models---including models trained adaptively against the users' past attacks, or models that use new technologies discovered after the attack.

We evaluate two systems for poisoning attacks against large-scale facial recognition, Fawkes (500,000+ downloads) and LowKey. We demonstrate how an "oblivious" model trainer can simply wait for future developments in computer vision to nullify the protection of pictures collected in the past. We further show that an adversary with black-box access to the attack can (i) train a robust model that resists the perturbations of collected pictures and (ii) detect poisoned pictures uploaded online.

We caution that facial recognition poisoning will not admit an "arms race" between attackers and defenders. Once perturbed pictures are scraped, the attack cannot be changed so any future successful defense irrevocably undermines users' privacy.

MetaMorph: Learning Universal Controllers with Transformers

- Agrim Gupta, Linxi Fan, Surya Ganguli, Li Fei-Fei

- abstract@open-review(Poster): Multiple domains like vision, natural language, and audio are witnessing tremendous progress by leveraging Transformers for large scale pre-training followed by task specific fine tuning. In contrast, in robotics we primarily train a single robot for a single task. However, modular robot systems now allow for the flexible combination of general-purpose building blocks into task optimized morphologies. However, given the exponentially large number of possible robot morphologies, training a controller for each new design is impractical. In this work, we propose MetaMorph, a Transformer based approach to learn a universal controller over a modular robot design space. MetaMorph is based on the insight that robot morphology is just another modality on which we can condition the output of a Transformer. Through extensive experiments we demonstrate that large scale pre-training on a variety of robot morphologies results in policies with combinatorial generalization capabilities, including zero shot generalization to unseen robot morphologies. We further demonstrate that our pre-trained policy can be used for sample-efficient transfer to completely new robot morphologies and tasks.

HTLM: Hyper-Text Pre-Training and Prompting of Language Models

- Armen Aghajanyan, Dmytro Okhonko, Mike Lewis, Mandar Joshi, Hu Xu, Gargi Ghosh, Luke Zettlemoyer
- abstract@open-review(Poster): We introduce HTLM, a hyper-text language model trained on a large-scale web crawl. Modeling hyper-text has a number of advantages: (1) it is easily gathered at scale, (2) it provides rich document-level and end-task-adjacent supervision (e.g. 'class' and 'id' attributes often encode document category information), and (3) it allows for new structured prompting that follows the established semantics of HTML (e.g. to do zero-shot summarization by infilling 'title' tags for a webpage that contains the input text). We show that pretraining with a BART-style denoising loss directly on simplified HTML provides highly effective transfer for a wide range of end tasks and supervision levels. HTLM matches or exceeds the performance of comparably sized text-only LMs for zero-shot prompting and fine-tuning for classification benchmarks, while also setting new state-of-the-art performance levels for zero-shot summarization. We also find that hyper-text prompts provide more value to HTLM, in terms of data efficiency, than plain text prompts do for existing LMs, and that HTLM is highly effective at auto-prompting itself, by simply generating the most likely hyper-text formatting for any available training data. We will release all code and models to support future HTLM research.

Illiterate DALL-E Learns to Compose

- Gautam Singh, Fei Deng, Sungjin Ahn
- abstract@open-review(Poster): Although DALL-E has shown an impressive ability of composition-based systematic generalization in image generation, it requires the dataset of text-image pairs and the compositionality is provided by the text. In contrast, object-centric representation models like the Slot Attention model learn composable representations without the text prompt. However, unlike DALL-E, its ability to systematically generalize for zero-shot generation is significantly limited. In this paper, we propose a simple but novel slot-based autoencoding architecture, called SLATE, for combining the best of both worlds: learning object-centric representations that allow systematic generalization in zero-shot image generation without text. As such, this model can also be seen as an illiterate DALL-E model. Unlike the pixel-mixture decoders of existing object-centric representation models, we propose to use the Image GPT decoder conditioned on the slots for capturing complex interactions among the slots and pixels. In experiments, we show that this simple and easy-to-implement architecture not requiring a text prompt achieves significant improvement in in-distribution and out-of-distribution (zero-shot) image generation and qualitatively comparable or better slot-attention structure than the models based on mixture decoders.

The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models

- Alexander Pan, Kush Bhatia, Jacob Steinhardt
- abstract@open-review(Poster): Reward hacking---where RL agents exploit gaps in misspecified proxy rewards---has been widely observed, but not yet systematically studied. To understand reward hacking, we construct four RL environments with different misspecified rewards. We investigate reward hacking as a function of agent capabilities: model capacity, action space resolution, and observation space noise. Typically, more capable agents are able to better exploit reward misspecifications, causing them to attain higher proxy reward and lower true reward. Moreover, we find instances of \emph{phase transitions}: capability thresholds at which the agent's behavior qualitatively shifts, leading to a sharp decrease in the true reward. Such phase transitions pose challenges to monitoring the safety of ML systems. To encourage further research on reward misspecification, address this, we propose an anomaly detection task for aberrant policies and offer several baseline detectors.

Optimizing Neural Networks with Gradient Lexicase Selection

- Li Ding, Lee Spector
- abstract@open-review(Poster): One potential drawback of using aggregated performance measurement in machine learning is that models may learn to accept higher errors on some training cases as compromises for lower errors on others, with the lower errors actually being instances of overfitting. This can lead both to stagnation at local optima and to poor generalization. Lexicase selection is an uncompromising method developed in evolutionary computation, which selects models on the basis of sequences of individual training case errors instead of using aggregated metrics such as loss and accuracy. In this paper, we investigate how the general idea of lexicase selection can fit into the context of deep learning to improve generalization. We propose Gradient Lexicase Selection, an optimization framework that combines gradient descent and lexicase selection in an evolutionary fashion. Experimental results show that the proposed method improves the generalization performance of various popular deep neural network architectures on three image classification benchmarks. Qualitative analysis also indicates that our method helps the networks learn more diverse representations.

Offline Reinforcement Learning with Implicit Q-Learning

- Ilya Kostrikov, Ashvin Nair, Sergey Levine
- abstract@open-review(Poster): Offline reinforcement learning requires reconciling two conflicting aims: learning a policy that improves over the behavior policy that collected the dataset, while at the same time minimizing the deviation from the behavior policy so as to avoid errors due to distributional shift. This tradeoff is critical, because most current offline reinforcement learning methods need to query the value of unseen actions during training to improve the policy, and therefore need to either constrain these actions to be in-distribution, or else regularize their values. We propose a new offline RL method that never needs to evaluate actions outside of the dataset, but still enables the learned policy to improve substantially over the best behavior in the data through generalization. The main insight in our work is that, instead of evaluating unseen actions from the latest policy, we can approximate the policy improvement step implicitly by treating the state value function as a random variable, with randomness determined by the action (while still integrating over the dynamics to avoid excessive optimism), and then taking a state conditional upper expectile of this random variable to estimate the value of the best actions in that state. This leverages the generalization capacity of the function approximator to estimate the value of the best available action at a given state without ever directly querying a Q-function with this unseen action. Our algorithm alternates between fitting this upper expectile value function and backing it up into a Q-function, without any explicit policy. Then, we extract the policy via advantage-weighted behavioral cloning, which also avoids querying out-of-sample actions. We dub our method Implicit Q-learning (IQL). IQL is easy to implement, computationally efficient, and only requires fitting an additional critic with an asymmetric L2 loss. IQL demonstrates the state-of-the-art performance on D4RL, a standard benchmark for offline reinforcement learning. We also demonstrate that IQL achieves strong performance fine-tuning using online interaction after offline initialization.

Learning Distributionally Robust Models at Scale via Composite Optimization

- Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, amin karbasi
- abstract@open-review(Poster): To train machine learning models that are robust to distribution shifts in the data, distributionally robust optimization (DRO) has been proven very effective. However, the existing approaches to learning a distributionally robust model either require solving complex

optimization problems such as semidefinite programming or a first-order method whose convergence scales linearly with the number of data samples--which hinders their scalability to large datasets. In this paper, we show how different variants of DRO are simply instances of a finite-sum composite optimization for which we provide scalable methods. We also provide empirical results that demonstrate the effectiveness of our proposed algorithm with respect to the prior art in order to learn robust models from very large datasets.

Counterfactual Plans under Distributional Ambiguity

- Ngoc Bui, Duy Nguyen, Viet Anh Nguyen
- abstract@open-review(Poster): Counterfactual explanations are attracting significant attention due to the flourishing applications of machine learning models in consequential domains. A counterfactual plan consists of multiple possibilities to modify a given instance so that the model's prediction will be altered. As the predictive model can be updated subject to the future arrival of new data, a counterfactual plan may become ineffective or infeasible, with respect to the future values of the model parameters. In this work, we study the counterfactual plans under model uncertainty, in which the distribution of the model parameters is partially prescribed using only the first- and second-moment information. First, we propose an uncertainty quantification tool to compute the lower and upper bounds of the probability of feasibility for any given counterfactual plan. We then provide corrective methods to adjust the counterfactual plan to improve the feasibility measure. The numerical experiments validate our bounds and demonstrate that our correction increases the robustness of the counterfactual plans in different real-world datasets.

Neural Parameter Allocation Search

- Bryan A. Plummer, Nikoli Dryden, Julius Frost, Torsten Hoefler, Kate Saenko
- abstract@open-review(Poster): Training neural networks requires increasing amounts of memory. Parameter sharing can reduce memory and communication costs, but existing methods assume networks have many identical layers and utilize hand-crafted sharing strategies that fail to generalize. We introduce Neural Parameter Allocation Search (NPAS), a novel task where the goal is to train a neural network given an arbitrary, fixed parameter budget. NPAS covers both low-budget regimes, which produce compact networks, as well as a novel high-budget regime, where additional capacity can be added to boost performance without increasing inference FLOPs. To address NPAS, we introduce Shapeshifter Networks (SSNs), which automatically learn where and how to share parameters in a network to support any parameter budget without requiring any changes to the architecture or loss function. NPAS and SSNs provide a complete framework for addressing generalized parameter sharing, and can also be combined with prior work for additional performance gains. We demonstrate the effectiveness of our approach using nine network architectures across four diverse tasks, including ImageNet classification and transformers.

Non-Linear Operator Approximations for Initial Value Problems

- Gaurav Gupta, Xiongye Xiao, Radu Balan, Paul Bogdan
- abstract@open-review(Poster): Time-evolution of partial differential equations is the key to model several dynamical processes, events forecasting but the operators associated with such problems are non-linear. We propose a PadÃ© approximation based exponential neural operator scheme for efficiently learning the map between a given initial condition and activities at a later time. The multiwavelets bases are used for space discretization. By explicitly embedding the exponential operators in the model, we reduce the training parameters and make it more data-efficient which is essential in dealing with scarce real-world datasets. The PadÃ© exponential operator uses a $\textit{recurrent structure with shared parameters}$ to model the non-linearity compared to recent neural operators that rely on using multiple linear operator layers in succession. We show theoretically that the gradients associated with the recurrent PadÃ© network are bounded across the recurrent horizon. We perform experiments on non-linear systems such as Korteweg-de Vries (KdV) and Kuramotoâ€“Sivashinsky (KS) equations to show that the proposed approach achieves the best performance and at the same time is data-efficient. We also show that urgent real-world problems like Epidemic forecasting (for example, COVID-19) can be formulated as a 2D time-varying operator problem. The proposed PadÃ© exponential operators yield better prediction results ($\textbf{53\%} (\textbf{52\%})$ better MAE than best neural operator (non-neural operator deep learning model)) compared to state-of-the-art forecasting models.

Constructing a Good Behavior Basis for Transfer using Generalized Policy Updates

- Safa Alver, Doina Precup
- abstract@open-review(Poster): We study the problem of learning a good set of policies, so that when combined together, they can solve a wide variety of unseen reinforcement learning tasks with no or very little new data. Specifically, we consider the framework of generalized policy evaluation and improvement, in which the rewards for all tasks of interest are assumed to be expressible as a linear combination of a fixed set of features. We show theoretically that, under certain assumptions, having access to a specific set of diverse policies, which we call a set of independent policies, can allow for instantaneously achieving high-level performance on all possible downstream tasks which are typically more complex than the ones on which the agent was trained. Based on this theoretical analysis, we propose a simple algorithm that iteratively constructs this set of policies. In addition to empirically validating our theoretical results, we compare our approach with recently proposed diverse policy set construction methods and show that, while others fail, our approach is able to build a behavior basis that enables instantaneous transfer to all possible downstream tasks. We also show empirically that having access to a set of independent policies can better bootstrap the learning process on downstream tasks where the new reward function cannot be described as a linear combination of the features. Finally, we demonstrate how this policy set can be useful in a lifelong reinforcement learning setting.

Collapse by Conditioning: Training Class-conditional GANs with Limited Data

- Mohamad Shahbazi, Martin Danelljan, Danda Pani Paudel, Luc Van Gool
- abstract@open-review(Poster): Class-conditioning offers a direct means to control a Generative Adversarial Network (GAN) based on a discrete input variable. While necessary in many applications, the additional information provided by the class labels could even be expected to benefit the training of the GAN itself. On the contrary, we observe that class-conditioning causes mode collapse in limited data settings, where unconditional learning leads to satisfactory generative ability. Motivated by this observation, we propose a training strategy for class-conditional GANs (cGANs) that effectively prevents the observed mode-collapse by leveraging unconditional learning. Our training strategy starts with an unconditional GAN and gradually injects the class conditioning into the generator and the objective function. The proposed method for training cGANs with limited data results not only in stable training but also in generating high-quality images, thanks to the early-stage exploitation of the shared information across classes. We analyze the observed mode collapse problem in comprehensive experiments on four datasets. Our approach demonstrates outstanding results compared with state-of-the-art methods and established baselines. The code is available at https://github.com/mshahbazi72/transitional-cGAN

High Probability Bounds for a Class of Nonconvex Algorithms with AdaGrad Stepsize

- Ali Kavis, Kfir Yehuda Levy, Volkan Cevher
- abstract@open-review(Poster): In this paper, we propose a new, simplified high probability analysis of AdaGrad for smooth, non-convex problems. More specifically, we focus on a particular accelerated gradient (AGD) template (Lan, 2020), through which we recover the original AdaGrad and its variant with averaging, and prove a convergence rate of $\mathcal O (1/ \sqrt{T})$ with high probability without the knowledge of smoothness and variance. We use a particular version of Freedman's concentration bound for martingale difference sequences (Kakade & Tewari, 2008) which enables us to achieve the best-known dependence of $\log (1 / \delta)$ on the probability margin δ. We present our analysis in a modular way and obtain a complementary $\mathcal O (1 / T)$ convergence rate in the deterministic setting. To the best of our knowledge, this is the first high probability result for AdaGrad with a

truly adaptive scheme, i.e., completely oblivious to the knowledge of smoothness and uniform variance bound, which simultaneously has best-known dependence of $\log(1/ \delta)$. We further prove noise adaptation property of AdaGrad under additional noise assumptions.

Map Induction: Compositional spatial submap learning for efficient exploration in novel environments

- Sugandha Sharma, Aidan Curtis, Marta Kryven, Joshua B. Tenenbaum, Ila R Fiete
- abstract@open-review(Poster): Humans are expert explorers and foragers. Understanding the computational cognitive mechanisms that support this capability can advance the study of the human mind and enable more efficient exploration algorithms. We hypothesize that humans explore new environments by inferring the structure of unobserved spaces through re-use of spatial information collected from previously explored spaces. Taking inspiration from the neuroscience of repeating map fragments and ideas about program induction, we present a novel ``Map Induction'' framework, which involves the generation of novel map proposals for unseen environments based on compositions of already-seen spaces in a Hierarchical Bayesian framework. The model thus explicitly reasons about unseen spaces through a distribution of strong spatial priors. We introduce a new behavioral Map Induction Task (MIT) that involves foraging for rewards to compare human performance with state-of-the-art existing models and Map Induction. We show that Map Induction better predicts human behavior than the non-inductive baselines. We also show that Map Induction, when used to augment state-of-the-art approximate planning algorithms, improves their performance.

How Did the Model Change? Efficiently Assessing Machine Learning API Shifts

- Lingjiao Chen, Matei Zaharia, James Zou
- abstract@open-review(Poster): ML prediction APIs from providers like Amazon and Google have made it simple to use ML in applications. A challenge for users is that such APIs continuously change over time as the providers update models, and changes can happen silently without users knowing. It is thus important to monitor when and how much the MLAPIsâ€™ performance shifts. To provide detailed change assessment, we model MLAPI shifts as confusion matrix differences, and propose a principled algorithmic framework, MASA, to provably assess these shifts efficiently given a sample budget constraint.MASAemploys an upper-confidence bound based approach to adaptively determine on which data point to query the ML API to estimate shifts. Empirically, we observe significant ML API shifts from 2020 to 2021 among 12 out of 36 applications using commercial APIs from Google, Microsoft, Amazon, and other providers. These real-world shifts include both improvements and reductions in accuracy. Extensive experiments show that MASA can estimate such API shifts more accurately than standard approaches given the same budget

A Tale of Two Flows: Cooperative Learning of Langevin Flow and Normalizing Flow Toward Energy-Based Model

- Jianwen Xie, Yaxuan Zhu, Jun Li, Ping Li
- abstract@open-review(Poster): This paper studies the cooperative learning of two generative flow models, in which the two models are iteratively updated based on the jointly synthesized examples. The first flow model is a normalizing flow that transforms an initial simple density to a target density by applying a sequence of invertible transformations. The second flow model is a Langevin flow that runs finite steps of gradient-based MCMC toward an energy-based model. We start from proposing a generative framework that trains an energy-based model with a normalizing flow as an amortized sampler to initialize the MCMC chains of the energy-based model. In each learning iteration, we generate synthesized examples by using a normalizing flow initialization followed by a short-run Langevin flow revision toward the current energy-based model. Then we treat the synthesized examples as fair samples from the energy-based model and update the model parameters with the maximum likelihood learning gradient, while the normalizing flow directly learns from the synthesized examples by maximizing the tractable likelihood. Under the short-run non-mixing MCMC scenario, the estimation of the energy-based model is shown to follow the perturbation of maximum likelihood, and the short-run Langevin flow and the normalizing flow form a two-flow generator that we call CoopFlow. We provide an understating of the CoopFlow algorithm by information geometry and show that it is a valid generator as it converges to a moment matching estimator. We demonstrate that the trained CoopFlow is capable of synthesizing realistic images, reconstructing images, and interpolating between images.

Robust Learning Meets Generative Models: Can Proxy Distributions Improve Adversarial Robustness?

- Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, Prateek Mittal
- abstract@open-review(Poster): While additional training data improves the robustness of deep neural networks against adversarial examples, it presents the challenge of curating a large number of specific real-world samples. We circumvent this challenge by using additional data from proxy distributions learned by advanced generative models. We first seek to formally understand the transfer of robustness from classifiers trained on proxy distributions to the real data distribution. We prove that the difference between the robustness of a classifier on the two distributions is upper bounded by the conditional Wasserstein distance between them. Next we use proxy distributions to significantly improve the performance of adversarial training on five different datasets. For example, we improve robust accuracy by up to 7.5% and 6.7% in ℓ_{∞} and ℓ_2 threat model over baselines that are not using proxy distributions on the CIFAR-10 dataset. We also improve certified robust accuracy by 7.6% on the CIFAR-10 dataset. We further demonstrate that different generative models brings a disparate improvement in the performance in robust training. We propose a robust discrimination approach to characterize the impact and further provide a deeper understanding of why diffusion-based generative models are a better choice for proxy distribution than generative adversarial networks.

Chaos is a Ladder: A New Theoretical Understanding of Contrastive Learning via Augmentation Overlap

- Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, Zhouchen Lin
- abstract@open-review(Poster): Recently, contrastive learning has risen to be a promising approach for large-scale self-supervised learning. However, theoretical understanding of how it works is still unclear. In this paper, we propose a new guarantee on the downstream performance without resorting to the conditional independence assumption that is widely adopted in previous work but hardly holds in practice. Our new theory hinges on the insight that the support of different intra-class samples will become more overlapped under aggressive data augmentations, thus simply aligning the positive samples (augmented views of the same sample) could make contrastive learning cluster intra-class samples together. Based on this augmentation overlap perspective, theoretically, we obtain asymptotically closed bounds for downstream performance under weaker assumptions, and empirically, we propose an unsupervised model selection metric ARC that aligns well with downstream accuracy. Our theory suggests an alternative understanding of contrastive learning: the role of aligning positive samples is more like a surrogate task than an ultimate goal, and the overlapped augmented views (i.e., the chaos) create a ladder for contrastive learning to gradually learn class-separated representations. The code for computing ARC is available at https://github.com/zhangq327/ARC.

Language-biased image classification: evaluation based on semantic representations

- Yoann Lemesle, Masataka Sawayama, Guillermo Valle-Perez, Maxime Adolphe, HÃ©lÃ¨ne SauzÃ©on, Pierre-Yves Oudeyer
- abstract@open-review(Poster): Humans show language-biased image recognition for a word-embedded image, known as picture-word interference. Such interference depends on hierarchical semantic categories and reflects that human language processing highly interacts with visual processing. Similar to humans, recent artificial models jointly trained on texts and images, e.g., OpenAI CLIP, show language-biased image classification. Exploring whether the bias leads to interference similar to those observed in humans can contribute to understanding how much the model acquires hierarchical semantic representations from joint learning of language and vision. The present study introduces methodological tools from the cognitive science literature to assess the biases of artificial models. Specifically, we introduce a benchmark task to test whether words superimposed on images can distort the image classification across different category levels and, if it can, whether the perturbation is due to the shared semantic representation between language and

vision. Our dataset is a set of word-embedded images and consists of a mixture of natural image datasets and hierarchical word labels with superordinate/basic category levels. Using this benchmark test, we evaluate the CLIP model. We show that presenting words distorts the image classification by the model across different category levels, but the effect does not depend on the semantic relationship between images and embedded words. This suggests that the semantic word representation in the CLIP visual processing is not shared with the image representation, although the word representation strongly dominates for word-embedded images.

Robbing the Fed: Directly Obtaining Private Data in Federated Learning with Modified Models

- Liam H Fowl, Jonas Geiping, Wojciech Czaja, Micah Goldblum, Tom Goldstein
- abstract@open-review(Poster): Federated learning has quickly gained popularity with its promises of increased user privacy and efficiency. Previous works have shown that federated gradient updates contain information that can be used to approximately recover user data in some situations. These previous attacks on user privacy have been limited in scope and do not scale to gradient updates aggregated over even a handful of data points, leaving some to conclude that data privacy is still intact for realistic training regimes. In this work, we introduce a new threat model based on minimal but malicious modifications of the shared model architecture which enable the server to directly obtain a verbatim copy of user data from gradient updates without solving difficult inverse problems. Even user data aggregated over large batches â€“ where previous methods fail to extract meaningful content â€“ can be reconstructed by these minimally modified models.

Practical Conditional Neural Process Via Tractable Dependent Predictions

- Stratis Markou, James Requeima, Wessel Bruinsma, Anna Vaughan, Richard E Turner
- abstract@open-review(Poster): Conditional Neural Processes (CNPs; Garnelo et al., 2018a) are meta-learning models which leverage the flexibility of deep learning to produce well-calibrated predictions and naturally handle off-the-grid and missing data. CNPs scale to large datasets and train with ease. Due to these features, CNPs appear well-suited to tasks from environmental sciences or healthcare. Unfortunately, CNPs do not produce correlated predictions, making them fundamentally inappropriate for many estimation and decision making tasks. Predicting heat waves or floods, for example, requires modelling dependencies in temperature or precipitation over time and space. Existing approaches which model output dependencies, such as Neural Processes (NPs; Garnelo et al., 2018b) or the FullConvGNP (Bruinsma et al., 2021), are either complicated to train or prohibitively expensive. What is needed is an approach which provides dependent predictions, but is simple to train and computationally tractable. In this work, we present a new class of Neural Process models that make correlated predictions and support exact maximum likelihood training that is simple and scalable. We extend the proposed models by using invertible output transformations, to capture non-Gaussian output distributions. Our models can be used in downstream estimation tasks which require dependent function samples. By accounting for output dependencies, our models show improved predictive performance on a range of experiments with synthetic and real data.

Reward Uncertainty for Exploration in Preference-based Reinforcement Learning

- Xinran Liang, Katherine Shu, Kimin Lee, Pieter Abbeel
- abstract@open-review(Poster): Conveying complex objectives to reinforcement learning (RL) agents often requires meticulous reward engineering. Preference-based RL methods are able to learn a more flexible reward model based on human preferences by actively incorporating human feedback, i.e. teacher's preferences between two clips of behaviors. However, poor feedback-efficiency still remains as a problem in current preference-based RL algorithms, as tailored human feedback is very expensive. To handle this issue, previous methods have mainly focused on improving query selection and policy initialization. At the same time, recent exploration methods have proven to be a recipe for improving sample-efficiency in RL. We present an exploration method specifically for preference-based RL algorithms. Our main idea is to design an intrinsic reward by measuring the novelty based on learned reward. Specifically, we utilize disagreement across ensemble of learned reward models. Our intuition is that disagreement in learned reward model reflects uncertainty in tailored human feedback and could be useful for exploration. Our experiments show that reward uncertainty exploration improves both feedback- and sample-efficiency of preference-based RL algorithms on complex robot manipulation tasks from Meta-World benchmarks, compared with other existing exploration methods that measure the novelty of state visitation.

Decentralized Learning for Overparameterized Problems: A Multi-Agent Kernel Approximation Approach

- Prashant Khanduri, Haibo Yang, Mingyi Hong, Jia Liu, Hoi To Wai, Sijia Liu
- abstract@open-review(Poster): This work develops a novel framework for communication-efficient distributed learning where the models to be learned are overparameterized. We focus on a class of kernel learning problems (which includes the popular neural tangent kernel (NTK) learning as a special case) and propose a novel {\it multi-agent kernel approximation} technique that allows the agents to distributedly estimate the full kernel function, and subsequently perform decentralized optimization, without directly exchanging any local data or parameters. The proposed framework is a significant departure from the classical consensus-based approaches, because the agents do not exchange problem parameters, and no consensus is required. We analyze the optimization and the generalization performance of the proposed framework for the ℓ_2 loss. We show that with M agents and N total samples when certain generalized inner-product kernels (resp. the random features kernel) are used, each agent needs to communicate $\mathcal{O}\big({N^2}/{M}\big)$ bits (resp. $\mathcal{O}\big(N \sqrt{N}/M \big)$ real values) to achieve minimax optimal generalization performance. We validate the theoretical results on 90 UCI benchmarking datasets (with average data size $N \approx 1000$) and show that each agent needs to share a total of $200N/M$ bits (resp. $3N/M$ real values) to closely match the performance of the centralized algorithms, and these numbers are independent of parameter and feature dimensions.

Permutation-Based SGD: Is Random Optimal?

- Shashank Rajput, Kangwook Lee, Dimitris Papailiopoulos
- abstract@open-review(Poster): A recent line of ground-breaking results for permutation-based SGD has corroborated a widely observed phenomenon: random permutations offer faster convergence than with-replacement sampling. However, is random optimal? We show that this depends heavily on what functions we are optimizing, and the convergence gap between optimal and random permutations can vary from exponential to nonexistent. We first show that for 1-dimensional strongly convex functions, with smooth second derivatives, there exist optimal permutations that offer exponentially faster convergence compared to random. However, for general strongly convex functions, random permutations are optimal. Finally, we show that for quadratic, strongly-convex functions, there are easy-to-construct permutations that lead to accelerated convergence compared to random. Our results suggest that a general convergence characterization of optimal permutations cannot capture the nuances of individual function classes, and can mistakenly indicate that one cannot do much better than random.

Graph-less Neural Networks: Teaching Old MLPs New Tricks Via Distillation

- Shichang Zhang, Yozen Liu, Yizhou Sun, Neil Shah
- abstract@open-review(Poster): Graph Neural Networks (GNNs) are popular for graph machine learning and have shown great results on wide node classification tasks. Yet, they are less popular for practical deployments in the industry owing to their scalability challenges incurred by data dependency. Namely, GNN inference depends on neighbor nodes multiple hops away from the target, and fetching them burdens latency-constrained applications. Existing inference acceleration methods like pruning and quantization can speed up GNNs by reducing Multiplication-and-ACcumulation (MAC) operations, but the improvements are limited given the data dependency is not resolved. Conversely, multi-layer perceptrons (MLPs) have no graph dependency and infer much faster than GNNs, even though they are less accurate than GNNs for node classification in general. Motivated by these

complementary strengths and weaknesses, we bring GNNs and MLPs together via knowledge distillation (KD). Our work shows that the performance of MLPs can be improved by large margins with GNN KD. We call the distilled MLPs Graph-less Neural Networks (GLNNs) as they have no inference graph dependency. We show that GLNNs with competitive accuracy infer faster than GNNs by 146X-273X and faster than other acceleration methods by 14X-27X. Under a production setting involving both transductive and inductive predictions across 7 datasets, GLNN accuracies improve over stand-alone MLPs by 12.36% on average and match GNNs on 6/7 datasets. Comprehensive analysis shows when and why GLNNs can achieve competitive accuracies to GNNs and suggests GLNN as a handy choice for latency-constrained applications.

Relating transformers to models and neural representations of the hippocampal formation

- James C. R. Whittington, Joseph Warren, Tim E.J. Behrens
- abstract@open-review(Poster): Many deep neural network architectures loosely based on brain networks have recently been shown to replicate neural firing patterns observed in the brain. One of the most exciting and promising novel architectures, the Transformer neural network, was developed without the brain in mind. In this work, we show that transformers, when equipped with recurrent position encodings, replicate the precisely tuned spatial representations of the hippocampal formation; most notably place and grid cells. Furthermore, we show that this result is no surprise since it is closely related to current hippocampal models from neuroscience. We additionally show the transformer version offers dramatic performance gains over the neuroscience version. This work continues to bind computations of artificial and brain networks, offers a novel understanding of the hippocampal-cortical interaction, and suggests how wider cortical areas may perform complex tasks beyond current neuroscience models such as language comprehension.

How many degrees of freedom do we need to train deep networks: a loss landscape perspective

- Brett W Larsen, Stanislav Fort, Nic Becker, Surya Ganguli
- abstract@open-review(Poster): A variety of recent works, spanning pruning, lottery tickets, and training within random subspaces, have shown that deep neural networks can be trained using far fewer degrees of freedom than the total number of parameters. We analyze this phenomenon for random subspaces by first examining the success probability of hitting a training loss sublevel set when training within a random subspace of a given training dimensionality. We find a sharp phase transition in the success probability from 0 to 1 as the training dimension surpasses a threshold. This threshold training dimension increases as the desired final loss decreases, but decreases as the initial loss decreases. We then theoretically explain the origin of this phase transition, and its dependence on initialization and final desired loss, in terms of properties of the high-dimensional geometry of the loss landscape. In particular, we show via Gordon's escape theorem, that the training dimension plus the Gaussian width of the desired loss sublevel set, projected onto a unit sphere surrounding the initialization, must exceed the total number of parameters for the success probability to be large. In several architectures and datasets, we measure the threshold training dimension as a function of initialization and demonstrate that it is a small fraction of the total parameters, implying by our theory that successful training with so few dimensions is possible precisely because the Gaussian width of low loss sublevel sets is very large. Moreover, we compare this threshold training dimension to more sophisticated ways of reducing training degrees of freedom, including lottery tickets as well as a new, analogous method: lottery subspaces.

Is Importance Weighting Incompatible with Interpolating Classifiers?

- Ke Alexander Wang, Niladri Shekhar Chatterji, Saminul Haque, Tatsunori Hashimoto
- abstract@open-review(Poster): Importance weighting is a classic technique to handle distribution shifts. However, prior work has presented strong empirical and theoretical evidence demonstrating that importance weights can have little to no effect on overparameterized neural networks. \emph{Is importance weighting truly incompatible with the training of overparameterized neural networks?} Our paper answers this in the negative. We show that importance weighting fails not because of the overparameterization, but instead, as a result of using exponentially-tailed losses like the logistic or cross-entropy loss. As a remedy, we show that polynomially-tailed losses restore the effects of importance reweighting in correcting distribution shift in overparameterized models. We characterize the behavior of gradient descent on importance weighted polynomially-tailed losses with overparameterized linear models, and theoretically demonstrate the advantage of using polynomially-tailed losses in a label shift setting. Surprisingly, our theory shows that using weights that are obtained by exponentiating the classical unbiased importance weights can improve performance. Finally, we demonstrate the practical value of our analysis with neural network experiments on a subpopulation shift and a label shift dataset. When reweighted, our loss function can outperform reweighted cross-entropy by as much as 9\% in test accuracy. Our loss function also gives test accuracies comparable to, or even exceeding, well-tuned state-of-the-art methods for correcting distribution shifts.

Neural Models for Output-Space Invariance in Combinatorial Problems

- Yatin Nandwani, Vidit Jain, Mausam ., Parag Singla
- abstract@open-review(Poster): Recently many neural models have been proposed to solve combinatorial puzzles by implicitly learning underlying constraints using their solved instances, such as sudoku or graph coloring (GCP). One drawback of the proposed architectures, which are often based on Graph Neural Networks (GNN) (Zhou et al., 2020), is that they cannot generalize across the size of the output space from which variables are assigned a value, for example, set of colors in a GCP, or board-size in sudoku. We call the output space for the variables as â€˜value-setâ€™ . While many works have demonstrated generalization of GNNs across graph size, there has been no study on how to design a GNN for achieving value-set invariance for problems that come from the same domain. For example, learning to solve 16 x 16 sudoku after being trained on only 9 x 9 sudokus, or coloring a 7 colorable graph after training on 4 colorable graphs. In this work, we propose novel methods to extend GNN based architectures to achieve value-set invariance. Specifically, our model builds on recently proposed Recurrent Relational Networks (RRN) (Palm et al., 2018). Our first approach exploits the graph-size invariance of GNNs by converting a multi-class node classification problem into a binary node classification problem. Our second approach works directly with multiple classes by adding multiple nodes corresponding to the values in the value-set, and then connecting variable nodes to value nodes depending on the problem initialization. Our experimental evaluation on three different combinatorial problems demonstrates that both our models perform well on our novel problem, compared to a generic neural reasoner. Between two of our models, we observe an inherent trade-off: while the binarized model gives better performance when trained on smaller value-sets, multi-valued model is much more memory efficient, resulting in improved performance when trained on larger value-sets, where binarized model fails to train.

StyleNeRF: A Style-based 3D Aware Generator for High-resolution Image Synthesis

- Jiatao Gu, Lingjie Liu, Peng Wang, Christian Theobalt
- abstract@open-review(Poster): We propose StyleNeRF, a 3D-aware generative model for photo-realistic high-resolution image synthesis with high multi-view consistency, which can be trained on unstructured 2D images. Existing approaches either cannot synthesize high-resolution images with fine details or yield clearly noticeable 3D-inconsistent artifacts. In addition, many of them lack control on style attributes and explicit 3D camera poses. To address these issues, StyleNeRF integrates the neural radiance field (NeRF) into a style-based generator to tackle the aforementioned challenges, i.e., improving rendering efficiency and 3D consistency for high-resolution image generation. To address the first issue, we perform volume rendering only to produce a low-resolution feature map, and progressively apply upsampling in 2D. To mitigate the inconsistencies caused by 2D upsampling, we propose multiple designs including a better upsampler choice and a new regularization loss to enforce 3D consistency. With these designs, StyleNeRF is able to synthesize high-resolution images at interactive rates while preserving 3D consistency at high quality. StyleNeRF also enables control of camera poses and different levels of styles, which can generalize to unseen views. It also supports challenging tasks such as style mixing, inversion and simple semantic edits.

The Role of Pretrained Representations for the OOD Generalization of RL Agents

- Frederik TrÃ¤uble, Andrea Dittadi, Manuel Wuthrich, Felix Widmaier, Peter Vincent Gehler, Ole Winther, Francesco Locatello, Olivier Bachem, Bernhard SchÃ¶lkopf, Stefan Bauer
- abstract@open-review(Poster): Building sample-efficient agents that generalize out-of-distribution (OOD) in real-world settings remains a fundamental unsolved problem on the path towards achieving higher-level cognition. One particularly promising approach is to begin with low-dimensional, pretrained representations of our world, which should facilitate efficient downstream learning and generalization. By training 240 representations and over 10,000 reinforcement learning (RL) policies on a simulated robotic setup, we evaluate to what extent different properties of pretrained VAE-based representations affect the OOD generalization of downstream agents. We observe that many agents are surprisingly robust to realistic distribution shifts, including the challenging sim-to-real case. In addition, we find that the generalization performance of a simple downstream proxy task reliably predicts the generalization performance of our RL agents under a wide range of OOD settings. Such proxy tasks can thus be used to select pretrained representations that will lead to agents that generalize.

Enabling Arbitrary Translation Objectives with Adaptive Tree Search

- Wang Ling, Wojciech Stokowiec, Domenic Donato, Chris Dyer, Lei Yu, Laurent Sartran, Austin Matthews
- abstract@open-review(Poster): We introduce an adaptive tree search algorithm, which is a deterministic variant of Monte Carlo tree search, that can find high-scoring outputs under translation models that make no assumptions about the form or structure of the search objective. This algorithm enables the exploration of new kinds of models that are unencumbered by constraints imposed to make decoding tractable, such as autoregressivity or conditional independence assumptions. When applied to autoregressive models, our algorithm has different biases than beam search has, which enables a new analysis of the role of decoding bias in autoregressive models. Empirically, we show that our adaptive tree search algorithm finds outputs with substantially better model scores compared to beam search in autoregressive models, and compared to reranking techniques in models whose scores do not decompose additively with respect to the words in the output. We also characterise the correlation of several translation model objectives with respect to BLEU. We find that while some standard models are poorly calibrated and benefit from the beam search bias, other often more robust models (autoregressive models tuned to maximize expected automatic metric scores, the noisy channel model and a newly proposed objective) benefit from increasing amounts of search using our proposed decoder, whereas the beam search bias limits the improvements obtained from such objectives. Thus, we argue that as models improve, the improvements may be masked by over-reliance on beam search or reranking based methods.

Proof Artifact Co-Training for Theorem Proving with Language Models

- Jesse Michael Han, Jason Rute, Yuhuai Wu, Edward Ayers, Stanislas Polu
- abstract@open-review(Poster): Labeled data for imitation learning of theorem proving in large libraries of formalized mathematics is scarce as such libraries require years of concentrated effort by human specialists to be built. This is particularly challenging when applying large Transformer language models to tactic prediction, because the scaling of performance with respect to model size is quickly disrupted in the data-scarce, easily-overfitted regime. We propose PACT (Proof Artifact Co-Training), a general methodology for extracting abundant self-supervised data from kernel-level proof terms for joint training alongside the usual tactic prediction objective. We apply this methodology to Lean,an interactive proof assistant which hosts some of the most sophisticated formalized mathematics to date. We instrument Lean with a neural theorem prover driven by a Transformer language model and show that PACT improves theorem proving success rate on a held-out suite of test theorems from 32% to 48%.

Mirror Descent Policy Optimization

- Manan Tomar, Lior Shani, Yonathan Efroni, Mohammad Ghavamzadeh
- abstract@open-review(Poster): Mirror descent (MD), a well-known first-order method in constrained convex optimization, has recently been shown as an important tool to analyze trust-region algorithms in reinforcement learning (RL). However, there remains a considerable gap between such theoretically analyzed algorithms and the ones used in practice. Inspired by this, we propose an efficient RL algorithm, called {\em mirror descent policy optimization} (MDPO). MDPO iteratively updates the policy by {\em approximately} solving a trust-region problem, whose objective function consists of two terms: a linearization of the standard RL objective and a proximity term that restricts two consecutive policies to be close to each other. Each update performs this approximation by taking multiple gradient steps on this objective function. We derive {\em on-policy} and {\em off-policy} variants of MDPO, while emphasizing important design choices motivated by the existing theory of MD in RL. We highlight the connections between on-policy MDPO and two popular trust-region RL algorithms: TRPO and PPO, and show that explicitly enforcing the trust-region constraint is in fact {\em not} a necessity for high performance gains in TRPO. We then show how the popular soft actor-critic (SAC) algorithm can be derived by slight modifications of off-policy MDPO. Overall, MDPO is derived from the MD principles, offers a unified approach to viewing a number of popular RL algorithms, and performs better than or on-par with TRPO, PPO, and SAC in a number of continuous and discrete control tasks.

A Loss Curvature Perspective on Training Instabilities of Deep Learning Models

- Justin Gilmer, Behrooz Ghorbani, Ankush Garg, Sneha Kudugunta, Behnam Neyshabur, David Cardoze, George Edward Dahl, Zachary Nado, Orhan Firat
- abstract@open-review(Poster): In this work, we study the evolution of the loss Hessian across many classification tasks in order to understand the effect the curvature of the loss has on the training dynamics. Whereas prior work has focused on how different learning rates affect the loss Hessian observed during training, we also analyze the effects of model initialization, architectural choices, and common training heuristics such as gradient clipping and learning rate warmup. Our results demonstrate that successful model and hyperparameter choices allow the early optimization trajectory to either avoid---or navigate out of---regions of high curvature and into flatter regions that tolerate a higher learning rate. Our results suggest a unifying perspective on how disparate mitigation strategies for training instability ultimately address the same underlying failure mode of neural network optimization, namely poor conditioning. Inspired by the conditioning perspective, we show that learning rate warmup can improve training stability just as much as batch normalization, layer normalization, MetaInit, GradInit, and Fixup initialization.

Cross-Domain Imitation Learning via Optimal Transport

- Arnaud Fickinger, Samuel Cohen, Stuart Russell, Brandon Amos
- abstract@open-review(Poster): Cross-domain imitation learning studies how to leverage expert demonstrations of one agent to train an imitation agent with a different embodiment or morphology. Comparing trajectories and stationary distributions between the expert and imitation agents is challenging because they live on different systems that may not even have the same dimensionality. We propose Gromov-Wasserstein Imitation Learning (GWIL), a method for cross-domain imitation that uses the Gromov-Wasserstein distance to align and compare states between the different spaces of the agents. Our theory formally characterizes the scenarios where GWIL preserves optimality, revealing its possibilities and limitations. We demonstrate the effectiveness of GWIL in non-trivial continuous control domains ranging from simple rigid transformation of the expert domain to arbitrary transformation of the state-action space.

Large-Scale Representation Learning on Graphs via Bootstrapping

- Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Mehdi Azabou, Eva L Dyer, Remi Munos, Petar VeliÄ koviÄ‡, Michal Valko
- abstract@open-review(Poster): Self-supervised learning provides a promising path towards eliminating the need for costly label information in representation learning on graphs. However, to achieve state-of-the-art performance, methods often need large numbers of negative examples and rely on complex augmentations. This can be prohibitively expensive, especially for large graphs. To address these challenges, we introduce Bootstrapped Graph Latents (BGRL) - a graph representation learning method that learns by predicting alternative augmentations of the input. BGRL uses only simple

augmentations and alleviates the need for contrasting with negative examples, and thus is scalable by design. BGRL outperforms or matches prior methods on several established benchmarks, while achieving a 2-10x reduction in memory costs. Furthermore, we show that BGRL can be scaled up to extremely large graphs with hundreds of millions of nodes in the semi-supervised regime, achieving state-of-the-art performance and improving over supervised baselines where representations are shaped only through label information. In particular, our solution centered on BGRL constituted one of the winning entries to the Open Graph Benchmark -Large Scale Challenge at KDD Cup 2021, on a graph orders of magnitudes larger than all previously available benchmarks, thus demonstrating the scalability and effectiveness of our approach.

Robust and Scalable SDE Learning: A Functional Perspective

- Scott Alexander Cameron, Tyron Luke Cameron, Arnu Pretorius, Stephen J. Roberts
- abstract@open-review(Poster): Stochastic differential equations provide a rich class of flexible generative models, capable of describing a wide range of spatio-temporal processes. A host of recent work looks to learn data-representing SDEs, using neural networks and other flexible function approximators. Despite these advances, learning remains computationally expensive due to the sequential nature of SDE integrators. In this work, we propose an importance-sampling estimator for probabilities of observations of SDEs for the purposes of learning. Crucially, the approach we suggest does not rely on such integrators. The proposed method produces lower-variance gradient estimates compared to algorithms based on SDE integrators and has the added advantage of being embarrassingly parallelizable. This facilitates the effective use of large-scale parallel hardware for massive decreases in computation time.

Neural Processes with Stochastic Attention: Paying more attention to the context dataset

- Mingyu Kim, Kyeong Ryeol Go, Se-Young Yun
- abstract@open-review(Poster): Neural processes (NPs) aim to stochastically complete unseen data points based on a given context dataset. NPs essentially leverage a given dataset as a context representation to derive a suitable identifier for a novel task. To improve the prediction accuracy, many variants of NPs have investigated context embedding approaches that generally design novel network architectures and aggregation functions satisfying permutation invariant. In this work, we propose a stochastic attention mechanism for NPs to capture appropriate context information. From the perspective of information theory, we demonstrate that the proposed method encourages context embedding to be differentiated from a target dataset, allowing NPs to consider features in a target dataset and context embedding independently. We observe that the proposed method can appropriately capture context embedding even under noisy data sets and restricted task distributions, where typical NPs suffer from a lack of context embeddings. We empirically show that our approach substantially outperforms conventional NPs in various domains through 1D regression, predator-prey model, and image completion. Moreover, the proposed method is also validated by MovieLens-10k dataset, a real-world problem.

Evaluating Disentanglement of Structured Representations

- RaphaÃ«l Dang-Nhu
- abstract@open-review(Poster): We introduce the first metric for evaluating disentanglement at individual hierarchy levels of a structured latent representation. Applied to object-centric generative models, this offers a systematic, unified approach to evaluating (i) object separation between latent slots (ii) disentanglement of object properties inside individual slots (iii) disentanglement of intrinsic and extrinsic object properties. We theoretically show that our framework gives stronger guarantees of selecting a good model than previous disentanglement metrics. Experimentally, we demonstrate that viewing object compositionality as a disentanglement problem addresses several issues with prior visual metrics of object separation. As a core technical component, we present the first representation probing algorithm handling slot permutation invariance.

Geometric Transformers for Protein Interface Contact Prediction

- Alex Morehead, Chen Chen, Jianlin Cheng
- abstract@open-review(Poster): Computational methods for predicting the interface contacts between proteins come highly sought after for drug discovery as they can significantly advance the accuracy of alternative approaches, such as protein-protein docking, protein function analysis tools, and other computational methods for protein bioinformatics. In this work, we present the Geometric Transformer, a novel geometry-evolving graph transformer for rotation and translation-invariant protein interface contact prediction, packaged within DeepInteract, an end-to-end prediction pipeline. DeepInteract predicts partner-specific protein interface contacts (i.e., inter-protein residue-residue contacts) given the 3D tertiary structures of two proteins as input. In rigorous benchmarks, DeepInteract, on challenging protein complex targets from the 13th and 14th CASP-CAPRI experiments as well as Docking Benchmark 5, achieves 14% and 1.1% top L/5 precision (L: length of a protein unit in a complex), respectively. In doing so, DeepInteract, with the Geometric Transformer as its graph-based backbone, outperforms existing methods for interface contact prediction in addition to other graph-based neural network backbones compatible with DeepInteract, thereby validating the effectiveness of the Geometric Transformer for learning rich relational-geometric features for downstream tasks on 3D protein structures.

Diurnal or Nocturnal? Federated Learning of Multi-branch Networks from Periodically Shifting Distributions

- Chen Zhu, Zheng Xu, Mingqing Chen, Jakub KoneÄ nÃ½, Andrew Hard, Tom Goldstein
- abstract@open-review(Poster): Federated learning has been deployed to train machine learning models from decentralized client data on mobile devices in practice. The clients available for training are observed to have periodically shifting distributions changing with the time of day, which can cause instability in training and degrade the model performance. In this paper, instead of modeling the distribution shift with a block-cyclic pattern as previous works, we model it with a mixture of distributions that gradually shifts between daytime and nighttime modes, and find this intuitive model to better match the observations in practical federated learning systems. Furthermore, we propose to jointly train a clustering model and a multi-branch network to allocate lightweight specialized branches to clients from different modes. A temporal prior is used to significantly boost the training performance. Experiments for image classification on EMNIST and CIFAR datasets, and next word prediction on the Stack Overflow dataset show that the proposed algorithm can counter the effects of the distribution shift and significantly improve the final model performance.

IGLU: Efficient GCN Training via Lazy Updates

- S Deepak Narayanan, Aditya Sinha, Prateek Jain, Purushottam Kar, SUNDARARAJAN SELLAMANICKAM
- abstract@open-review(Poster): Training multi-layer Graph Convolution Networks (GCN) using standard SGD techniques scales poorly as each descent step ends up updating node embeddings for a large portion of the graph. Recent attempts to remedy this sub-sample the graph that reduces compute but introduce additional variance and may offer suboptimal performance. This paper develops the IGLU method that caches intermediate computations at various GCN layers thus enabling lazy updates that significantly reduce the compute cost of descent. IGLU introduces bounded bias into the gradients but nevertheless converges to a first-order saddle point under standard assumptions such as objective smoothness. Benchmark experiments show that IGLU offers up to 1.2% better accuracy despite requiring up to 88% less compute.

Procedural generalization by planning with self-supervised world models

- Ankesh Anand, Jacob C Walker, Yazhe Li, Eszter VÃ©rtes, Julian Schrittwieser, Sherjil Ozair, Theophane Weber, Jessica B Hamrick

- abstract@open-review(Poster): One of the key promises of model-based reinforcement learning is the ability to generalize using an internal model of the world to make predictions in novel environments and tasks. However, the generalization ability of model-based agents is not well understood because existing work has focused on model-free agents when benchmarking generalization. Here, we explicitly measure the generalization ability of model-based agents in comparison to their model-free counterparts. We focus our analysis on MuZero (Schrittwieser et al., 2020), a powerful model-based agent, and evaluate its performance on both procedural and task generalization. We identify three factors of procedural generalization---planning, self-supervised representation learning, and procedural data diversity---and show that by combining these techniques, we achieve state-of-the art generalization performance and data efficiency on Procgen (Cobbe et al., 2019). However, we find that these factors do not always provide the same benefits for the task generalization benchmarks in Meta-World (Yu et al., 2019), indicating that transfer remains a challenge and may require different approaches than procedural generalization. Overall, we suggest that building generalizable agents requires moving beyond the single-task, model-free paradigm and towards self-supervised model-based agents that are trained in rich, procedural, multi-task environments.

Top-N: Equivariant Set and Graph Generation without Exchangeability

- Clement Vignac, Pascal Frossard
- abstract@open-review(Poster): This work addresses one-shot set and graph generation, and, more specifically, the parametrization of probabilistic decoders that map a vector-shaped prior to a distribution over sets or graphs. Sets and graphs are most commonly generated by first sampling points i.i.d. from a normal distribution, and then processing these points along with the prior vector using Transformer layers or Graph Neural Networks. This architecture is designed to generate exchangeable distributions, i.e., all permutations of the generated outputs are equally likely. We however show that it only optimizes a proxy to the evidence lower bound, which makes it hard to train. We then study equivariance in generative settings and show that non-exchangeable methods can still achieve permutation equivariance. Using this result, we introduce Top-n creation, a differentiable generation mechanism that uses the latent vector to select the most relevant points from a trainable reference set. Top-n can replace i.i.d. generation in any Variational Autoencoder or Generative Adversarial Network. Experimentally, our method outperforms i.i.d. generation by 15% at SetMNIST reconstruction, by 33% at object detection on CLEVR, generates sets that are 74% closer to the true distribution on a synthetic molecule-like dataset, and generates more valid molecules on QM9.

The Spectral Bias of Polynomial Neural Networks

- Moulik Choraria, Leello Tadesse Dadi, Grigorios Chrysos, Julien Mairal, Volkan Cevher
- abstract@open-review(Poster): Polynomial neural networks (PNNs) have been recently shown to be particularly effective at image generation and face recognition, where high-frequency information is critical. Previous studies have revealed that neural networks demonstrate a $\text{\it{spectral bias}}$ towards low-frequency functions, which yields faster learning of low-frequency components during training. Inspired by such studies, we conduct a spectral analysis of the Neural Tangent Kernel (NTK) of PNNs. We find that the Π-Net family, i.e., a recently proposed parametrization of PNNs, speeds up the learning of the higher frequencies. We verify the theoretical bias through extensive experiments. We expect our analysis to provide novel insights into designing architectures and learning frameworks by incorporating multiplicative interactions via polynomials.

Invariant Causal Representation Learning for Out-of-Distribution Generalization

- Chaochao Lu, Yuhuai Wu, JosÃ© Miguel HernÃ¡ndez-Lobato, Bernhard SchÃ¶lkopf
- abstract@open-review(Poster): Due to spurious correlations, machine learning systems often fail to generalize to environments whose distributions differ from the ones used at training time. Prior work addressing this, either explicitly or implicitly, attempted to find a data representation that has an invariant relationship with the target. This is done by leveraging a diverse set of training environments to reduce the effect of spurious features and build an invariant predictor. However, these methods have generalization guarantees only when both data representation and classifiers come from a linear model class. We propose invariant Causal Representation Learning (iCaRL), an approach that enables out-of-distribution (OOD) generalization in the nonlinear setting (i.e., nonlinear representations and nonlinear classifiers). It builds upon a practical and general assumption: the prior over the data representation (i.e., a set of latent variables encoding the data) given the target and the environment belongs to general exponential family distributions, i.e., a more flexible conditionally non-factorized prior that can actually capture complicated dependences between the latent variables. Based on this, we show that it is possible to identify the data representation up to simple transformations. We also show that all direct causes of the target can be fully discovered, which further enables us to obtain generalization guarantees in the nonlinear setting. Experiments on both synthetic and real-world datasets demonstrate that our approach outperforms a variety of baseline methods.

LFPT5: A Unified Framework for Lifelong Few-shot Language Learning Based on Prompt Tuning of T5

- Chengwei Qin, Shafiq Joty
- abstract@open-review(Poster): Existing approaches to lifelong language learning rely on plenty of labeled data for learning a new task, which is hard to obtain in most real scenarios. Considering that humans can continually learn new tasks from a handful of examples, we expect the models also to be able to generalize well on new few-shot tasks without forgetting the previous ones. In this work, we define this more challenging yet practical problem as Lifelong Few-shot Language Learning (LFLL) and propose a unified framework for it based on prompt tuning of T5. Our framework called LFPT5 takes full advantage of PT's strong few-shot learning ability, and simultaneously trains the model as a task solver and a data generator. Before learning a new domain of the same task type, LFPT5 generates pseudo (labeled) samples of previously learned domains, and later gets trained on those samples to alleviate forgetting of previous knowledge as it learns the new domain. In addition, a KL divergence loss is minimized to achieve label consistency between the previous and the current model. While adapting to a new task type, LFPT5 includes and tunes additional prompt embeddings for the new task. With extensive experiments, we demonstrate that LFPT5 can be applied to various different types of tasks and significantly outperform previous methods in different LFLL settings.

On Non-Random Missing Labels in Semi-Supervised Learning

- Xinting Hu, Yulei Niu, Chunyan Miao, Xian-Sheng Hua, Hanwang Zhang
- abstract@open-review(Poster): Semi-Supervised Learning (SSL) is fundamentally a missing label problem, in which the label Missing Not At Random (MNAR) problem is more realistic and challenging, compared to the widely-adopted yet naive Missing Completely At Random assumption where both labeled and unlabeled data share the same class distribution. Different from existing SSL solutions that overlook the role of "class" in causing the non-randomness, e.g., users are more likely to label popular classes, we explicitly incorporate "class" into SSL. Our method is three-fold: 1) We propose Class-Aware Propensity (CAP) that exploits the unlabeled data to train an improved classifier using the biased labeled data. 2) To encourage rare class training, whose model is low-recall but high-precision that discards too many pseudo-labeled data, we propose Class-Aware Imputation (CAI) that dynamically decreases (or increases) the pseudo-label assignment threshold for rare (or frequent) classes. 3) Overall, we integrate CAP and CAI into a Class-Aware Doubly Robust (CADR) estimator for training an unbiased SSL model. Under various MNAR settings and ablations, our method not only significantly outperforms existing baselines, but also surpasses other label bias removal SSL methods.

Mapping conditional distributions for domain adaptation under generalized target shift

- Matthieu Kirchmeyer, Alain Rakotomamonjy, Emmanuel de Bezenac, patrick gallinari
- abstract@open-review(Poster): We consider the problem of unsupervised domain adaptation (UDA) between a source and a target domain under conditional and label shift a.k.a Generalized Target Shift (GeTarS). Unlike simpler UDA settings, few works have addressed this challenging problem.

Recent approaches learn domain-invariant representations, yet they have practical limitations and rely on strong assumptions that may not hold in practice. In this paper, we explore a novel and general approach to align pretrained representations, which circumvents existing drawbacks. Instead of constraining representation invariance, it learns an optimal transport map, implemented as a NN, which maps source representations onto target ones. Our approach is flexible and scalable, it preserves the problem's structure and it has strong theoretical guarantees under mild assumptions. In particular, our solution is unique, matches conditional distributions across domains, recovers target proportions and explicitly controls the target generalization risk. Through an exhaustive comparison on several datasets, we challenge the state-of-the-art in GeTarS.

On the Generalization of Models Trained with SGD: Information-Theoretic Bounds and Implications

- Ziqiao Wang, Yongyi Mao
- abstract@open-review(Poster): This paper follows up on a recent work of Neu et al. (2021) and presents some new information-theoretic upper bounds for the generalization error of machine learning models, such as neural networks, trained with SGD. We apply these bounds to analyzing the generalization behaviour of linear and two-layer ReLU networks. Experimental study of these bounds provide some insights on the SGD training of neural networks. They also point to a new and simple regularization scheme which we show performs comparably to the current state of the art.

Amortized Implicit Differentiation for Stochastic Bilevel Optimization

- Michael Arbel, Julien Mairal
- abstract@open-review(Poster): We study a class of algorithms for solving bilevel optimization problems in both stochastic and deterministic settings when the inner-level objective is strongly convex. Specifically, we consider algorithms based on inexact implicit differentiation and we exploit a warm-start strategy to amortize the estimation of the exact gradient. We then introduce a unified theoretical framework inspired by the study of singularly perturbed systems to analyze such amortized algorithms. By using this framework, our analysis shows these algorithms to match the computational complexity of oracle methods that have access to an unbiased estimate of the gradient, thus outperforming many existing results for bilevel optimization. We illustrate these findings on synthetic experiments and demonstrate the efficiency of these algorithms on hyper-parameter optimization experiments involving several thousands of variables.

Multi-objective Optimization by Learning Space Partition

- Yiyang Zhao, Linnan Wang, Kevin Yang, Tianjun Zhang, Tian Guo, Yuandong Tian
- abstract@open-review(Poster): In contrast to single-objective optimization (SOO), multi-objective optimization (MOO) requires an optimizer to find the Pareto frontier, a subset of feasible solutions that are not dominated by other feasible solutions. In this paper, we propose LaMOO, a novel multi-objective optimizer that learns a model from observed samples to partition the search space and then focus on promising regions that are likely to contain a subset of the Pareto frontier. The partitioning is based on the dominance number, which measures "how close" a data point is to the Pareto frontier among existing samples. To account for possible partition errors due to limited samples and model mismatch, we leverage Monte Carlo Tree Search (MCTS) to exploit promising regions while exploring suboptimal regions that may turn out to contain good solutions later. Theoretically, we prove the efficacy of learning space partitioning via LaMOO under certain assumptions. Empirically, on the HyperVolume (HV) benchmark, a popular MOO metric, LaMOO substantially outperforms strong baselines on multiple real-world MOO tasks, by up to 225% in sample efficiency for neural architecture search on Nasbench201, and up to 10% for molecular design.

Mapping Language Models to Grounded Conceptual Spaces

- Roma Patel, Ellie Pavlick
- abstract@open-review(Poster): A fundamental criticism of text-only language models (LMs) is their lack of grounding---that is, the ability to tie a word for which they have learned a representation, to its actual use in the world. However, despite this limitation, large pre-trained LMs have been shown to have a remarkable grasp of the conceptual structure of language, as demonstrated by their ability to answer questions, generate fluent text, or make inferences about entities, objects, and properties that they have never physically observed. In this work we investigate the extent to which the rich conceptual structure that LMs learn indeed reflects the conceptual structure of the non-linguistic world---which is something that LMs have never observed. We do this by testing whether the LMs can learn to map an entire conceptual domain (e.g., direction or colour) onto a grounded world representation given only a small number of examples. For example, we show a model what the word `left" means using a textual depiction of a grid world, and assess how well it can generalise to related concepts, for example, the word`right", in a similar grid world. We investigate a range of generative language models of varying sizes (including GPT-2 and GPT-3), and see that although the smaller models struggle to perform this mapping, the largest model can not only learn to ground the concepts that it is explicitly taught, but appears to generalise to several instances of unseen concepts as well. Our results suggest an alternative means of building grounded language models: rather than learning grounded representations ``from scratch", it is possible that large text-only models learn a sufficiently rich conceptual structure that could allow them to be grounded in a data-efficient way.

The Efficiency Misnomer

- Mostafa Dehghani, Yi Tay, Anurag Arnab, Lucas Beyer, Ashish Vaswani
- abstract@open-review(Poster): Model efficiency is a critical aspect of developing and deploying machine learning models. Inference time and latency directly affect the user experience, and some applications have hard requirements. In addition to inference costs, model training also have direct financial and environmental impacts. Although there are numerous well-established metrics (cost indicators) for measuring model efficiency, researchers and practitioners often assume that these metrics are correlated with each other and report only a few of them. In this paper, we thoroughly discuss common cost indicators, their advantages and disadvantages, and how they can contradict each other. We demonstrate how incomplete reporting of cost indicators can lead to partial conclusions and a blurred or incomplete picture of the practical considerations of different models. We further present suggestions to improve reporting of efficiency metrics.

Hybrid Memoised Wake-Sleep: Approximate Inference at the Discrete-Continuous Interface

- Tuan Anh Le, Katherine M. Collins, Luke Hewitt, Kevin Ellis, Siddharth N, Samuel Gershman, Joshua B. Tenenbaum
- abstract@open-review(Poster): Modeling complex phenomena typically involves the use of both discrete and continuous variables. Such a setting applies across a wide range of problems, from identifying trends in time-series data to performing effective compositional scene understanding in images. Here, we propose Hybrid Memoised Wake-Sleep (HMWS), an algorithm for effective inference in such hybrid discrete-continuous models. Prior approaches to learning suffer as they need to perform repeated expensive inner-loop discrete inference. We build on a recent approach, Memoised Wake-Sleep (MWS), which alleviates part of the problem by memoising discrete variables, and extend it to allow for a principled and effective way to handle continuous variables by learning a separate recognition model used for importance-sampling based approximate inference and marginalization. We evaluate HMWS in the GP-kernel learning and 3D scene understanding domains, and show that it outperforms current state-of-the-art inference methods.

Adversarial Retriever-Ranker for Dense Text Retrieval

- Hang Zhang, Yeyun Gong, Yelong Shen, Jiancheng Lv, Nan Duan, Weizhu Chen

- abstract@open-review(Poster): Current dense text retrieval models face two typical challenges. First, it adopts a siamese dual-encoder architecture to encode query and document independently for fast indexing and searching, whereas neglecting the finer-grained term-wise interactions. This results in a sub-optimal recall performance. Second, it highly relies on a negative sampling technique to build up the negative documents in its contrastive loss. To address these challenges, we present Adversarial Retriever-Ranker (AR2), which consists of a dual-encoder retriever plus a cross-encoder ranker. The two models are jointly optimized according to a minimax adversarial objective: the retriever learns to retrieve negative documents to cheat the ranker, while the ranker learns to rank a collection of candidates including both the ground-truth and the retrieved ones, as well as providing progressive direct feedback to the dual-encoder retriever. Through this adversarial game, the retriever gradually produces harder negative documents to train a better ranker, whereas the cross-encoder ranker provides progressive feedback to improve retriever. We evaluate AR2 on three benchmarks. Experimental results show that AR2 consistently and significantly outperforms existing dense retriever methods and achieves new state-of-the-art results on all of them. This includes the improvements on Natural Questions R@5 to 77.9%(+2.1%), TriviaQA R@5 to 78.2%(+1.4), and MS-MARCO MRR@10 to 39.5%(+1.3%). We will make our code, models, and data publicly available.

Conditioning Sequence-to-sequence Networks with Learned Activations

- Alberto Gil Couto Pimentel Ramos, Abhinav Mehrotra, Nicholas Donald Lane, Sourav Bhattacharya
- abstract@open-review(Poster): Conditional neural networks play an important role in a number of sequence-to-sequence modeling tasks, including personalized sound enhancement (PSE), speaker dependent automatic speech recognition (ASR), and generative modeling such as text-to-speech synthesis. In conditional neural networks, the output of a model is often influenced by a conditioning vector, in addition to the input. Common approaches of conditioning include input concatenation or modulation with the conditioning vector, which comes at a cost of increased model size. In this work, we introduce a novel approach of neural network conditioning by learning intermediate layer activations based on the conditioning vector. We systematically explore and show that learned activation functions can produce conditional models with comparable or better quality, while decreasing model sizes, thus making them ideal candidates for resource-efficient on-device deployment. As exemplary target use-cases we consider (i) the task of PSE as a pre-processing technique for improving telephony or pre-trained ASR performance under noise, and (ii) personalized ASR in single speaker scenarios. We find that conditioning via activation function learning is an effective modeling strategy, suggesting a broad applicability of the proposed technique across a number of application domains.

LOSSY COMPRESSION WITH DISTRIBUTION SHIFT AS ENTROPY CONSTRAINED OPTIMAL TRANSPORT

- Huan Liu, George Zhang, Jun Chen, Ashish J Khisti
- abstract@open-review(Poster): We study an extension of lossy compression where the reconstruction distribution is different from the source distribution in order to account for distributional shift due to processing. We formulate this as a generalization of optimal transport with an entropy bottleneck to account for the rate constraint due to compression. We provide expressions for the tradeoff between compression rate and the achievable distortion with and without shared common randomness between the encoder and decoder. We study the examples of binary, uniform and Gaussian sources (in an asymptotic setting) in detail and demonstrate that shared randomness can strictly improve the tradeoff. For the case without common randomness and squared-Euclidean distortion, we show that the optimal solution partially decouples into the problem of optimal compression and transport and also characterize the penalty associated with fully decoupling them. We provide experimental results by training deep learning end-to-end compression systems for performing denoising on SVHN and super-resolution on MNIST suggesting consistency with our theoretical results.

Equivariant Self-Supervised Learning: Encouraging Equivariance in Representations

- Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, Marin Soljacic
- abstract@open-review(Poster): In state-of-the-art self-supervised learning (SSL) pre-training produces semantically good representations by encouraging them to be invariant under meaningful transformations prescribed from human knowledge. In fact, the property of invariance is a trivial instance of a broader class called equivariance, which can be intuitively understood as the property that representations transform according to the way the inputs transform. Here, we show that rather than using only invariance, pre-training that encourages non-trivial equivariance to some transformations, while maintaining invariance to other transformations, can be used to improve the semantic quality of representations. Specifically, we extend popular SSL methods to a more general framework which we name Equivariant Self-Supervised Learning (E-SSL). In E-SSL, a simple additional pre-training objective encourages equivariance by predicting the transformations applied to the input. We demonstrate E-SSLâ€™s effectiveness empirically on several popular computer vision benchmarks, e.g. improving SimCLR to 72.5% linear probe accuracy on ImageNet. Furthermore, we demonstrate usefulness of E-SSL for applications beyond computer vision; in particular, we show its utility on regression problems in photonics science. Our code, datasets and pre-trained models are available at https://github.com/rdangovs/essl to aid further research in E-SSL.

Direct then Diffuse: Incremental Unsupervised Skill Discovery for State Covering and Goal Reaching

- Pierre-Alexandre Kamienny, Jean Tarbouriech, sylvain lamprier, Alessandro Lazaric, Ludovic Denoyer
- abstract@open-review(Poster): Learning meaningful behaviors in the absence of reward is a difficult problem in reinforcement learning. A desirable and challenging unsupervised objective is to learn a set of diverse skills that provide a thorough coverage of the state space while being directed, i.e., reliably reaching distinct regions of the environment. In this paper, we build on the mutual information framework for skill discovery and introduce UPSIDE, which addresses the coverage-directedness trade-off in the following ways: 1) We design policies with a decoupled structure of a directed skill, trained to reach a specific region, followed by a diffusing part that induces a local coverage. 2) We optimize policies by maximizing their number under the constraint that each of them reaches distinct regions of the environment (i.e., they are sufficiently discriminable) and prove that this serves as a lower bound to the original mutual information objective. 3) Finally, we compose the learned directed skills into a growing tree that adaptively covers the environment. We illustrate in several navigation and control environments how the skills learned by UPSIDE solve sparse-reward downstream tasks better than existing baselines.

Normalization of Language Embeddings for Cross-Lingual Alignment

- Prince Osei Aboagye, Yan Zheng, Chin-Chia Michael Yeh, Junpeng Wang, Wei Zhang, Liang Wang, Hao Yang, Jeff Phillips
- abstract@open-review(Poster): Learning a good transfer function to map the word vectors from two languages into a shared cross-lingual word vector space plays a crucial role in cross-lingual NLP. It is useful in translation tasks and important in allowing complex models built on a high-resource language like English to be directly applied on an aligned low resource language. While Procrustes and other techniques can align language models with some success, it has recently been identified that structural differences (for instance, due to differing word frequency) create different profiles for various monolingual embedding. When these profiles differ across languages, it correlates with how well languages can align and their performance on cross-lingual downstream tasks. In this work, we develop a very general language embedding normalization procedure, building and subsuming various previous approaches, which removes these structural profiles across languages without destroying their intrinsic meaning. We demonstrate that meaning is retained and alignment is improved on similarity, translation, and cross-language classification tasks. Our proposed normalization clearly outperforms all prior approaches like centering and vector normalization on each task and with each alignment approach.

Boosting the Certified Robustness of L-infinity Distance Nets

- Bohang Zhang, Du Jiang, Di He, Liwei Wang

- abstract@open-review(Poster): Recently, Zhang et al. (2021) developed a new neural network architecture based on ℓ_∞-distance functions, which naturally possesses certified ℓ_∞ robustness by its construction. Despite the novel design and theoretical foundation, so far the model only achieved comparable performance to conventional networks. In this paper, we make the following two contributions: $\mathrm{(i)}$ We demonstrate that ℓ_∞-distance nets enjoy a fundamental advantage in certified robustness over conventional networks (under typical certification approaches); $\mathrm{(ii)}$ With an improved training process we are able to significantly boost the certified accuracy of ℓ_∞-distance nets. Our training approach largely alleviates the optimization problem that arose in the previous training scheme, in particular, the unexpected large Lipschitz constant due to the use of a crucial trick called \textit{ℓ_p-relaxation}. The core of our training approach is a novel objective function that combines scaled cross-entropy loss and clipped hinge loss with a decaying mixing coefficient. Experiments show that using the proposed training strategy, the certified accuracy of ℓ_∞-distance net can be dramatically improved from 33.30% to 40.06% on CIFAR-10 ($\epsilon=8/255$), meanwhile outperforming other approaches in this area by a large margin. Our results clearly demonstrate the effectiveness and potential of ℓ_∞-distance net for certified robustness. Codes are available at https://github.com/zbh2047/L_inf-dist-net-v2.

Stochastic Training is Not Necessary for Generalization

- Jonas Geiping, Micah Goldblum, Phil Pope, Michael Moeller, Tom Goldstein
- abstract@open-review(Poster): It is widely believed that the implicit regularization of SGD is fundamental to the impressive generalization behavior we observe in neural networks. In this work, we demonstrate that non-stochastic full-batch training can achieve comparably strong performance to SGD on CIFAR-10 using modern architectures. To this end, we show that the implicit regularization of SGD can be completely replaced with explicit regularization. Our observations indicate that the perceived difficulty of full-batch training may be the result of its optimization properties and the disproportionate time and effort spent by the ML community tuning optimizers and hyperparameters for small-batch training.

Transfer RL across Observation Feature Spaces via Model-Based Regularization

- Yanchao Sun, Ruijie Zheng, Xiyao Wang, Andrew E Cohen, Furong Huang
- abstract@open-review(Poster): In many reinforcement learning (RL) applications, the observation space is specified by human developers and restricted by physical realizations, and may thus be subject to dramatic changes over time (e.g. increased number of observable features). However, when the observation space changes, the previous policy will likely fail due to the mismatch of input features, and another policy must be trained from scratch, which is inefficient in terms of computation and sample complexity. Following theoretical insights, we propose a novel algorithm which extracts the latent-space dynamics in the source task, and transfers the dynamics model to the target task to use as a model-based regularizer. Our algorithm works for drastic changes of observation space (e.g. from vector-based observation to image-based observation), without any inter-task mapping or any prior knowledge of the target task. Empirical results show that our algorithm significantly improves the efficiency and stability of learning in the target task.

GATSBI: Generative Adversarial Training for Simulation-Based Inference

- Poornima Ramesh, Jan-Matthis Lueckmann, Jan Boelts, Ã lvaro Tejero-Cantero, David S. Greenberg, Pedro J. Goncalves, Jakob H. Macke
- abstract@open-review(Poster): Simulation-based inference (SBI) refers to statistical inference on stochastic models for which we can generate samples, but not compute likelihoods. Like SBI algorithms, generative adversarial networks (GANs) do not require explicit likelihoods. We study the relationship between SBI and GANs, and introduce GATSBI, an adversarial approach to SBI. GATSBI reformulates the variational objective in an adversarial setting to learn implicit posterior distributions. Inference with GATSBI is amortised across observations, works in high-dimensional posterior spaces and supports implicit priors. We evaluate GATSBI on two common SBI benchmark problems and on two high-dimensional simulators. On a model for wave propagation on the surface of a shallow water body, we show that GATSBI can return well-calibrated posterior estimates even in high dimensions. On a model of camera optics, it infers a high-dimensional posterior given an implicit prior, and performs better than a state-of-the-art SBI approach. We also show how GATSBI can be extended to perform sequential posterior estimation to focus on individual observations. Overall, GATSBI opens up opportunities for leveraging advances in GANs to perform Bayesian inference on high-dimensional simulation-based models.

Domain Adversarial Training: A Game Perspective

- David Acuna, Marc T Law, Guojun Zhang, Sanja Fidler
- abstract@open-review(Poster): The dominant line of work in domain adaptation has focused on learning invariant representations using domain-adversarial training. In this paper, we interpret this approach from a game theoretical perspective. Defining optimal solutions in domain-adversarial training as a local Nash equilibrium, we show that gradient descent in domain-adversarial training can violate the asymptotic convergence guarantees of the optimizer, oftentimes hindering the transfer performance. Our analysis leads us to replace gradient descent with high-order ODE solvers (i.e., Rungeâ€“Kutta), for which we derive asymptotic convergence guarantees. This family of optimizers is significantly more stable and allows more aggressive learning rates, leading to high performance gains when used as a drop-in replacement over standard optimizers. Our experiments show that in conjunction with state-of-the-art domain-adversarial methods, we achieve up to 3.5% improvement with less than of half training iterations. Our optimizers are easy to implement, free of additional parameters, and can be plugged into any domain-adversarial framework.

Differentiable Expectation-Maximization for Set Representation Learning

- Minyoung Kim
- abstract@open-review(Poster): We tackle the set2vec problem, the task of extracting a vector representation from an input set comprised of a variable number of feature vectors. Although recent approaches based on self attention such as (Set)Transformers were very successful due to the capability of capturing complex interaction between set elements, the computational overhead is the well-known downside. The inducing-point attention and the latest optimal transport kernel embedding (OTKE) are promising remedies that attain comparable or better performance with reduced computational cost, by incorporating a fixed number of learnable queries in attention. In this paper we approach the set2vec problem from a completely different perspective. The elements of an input set are considered as i.i.d.~samples from a mixture distribution, and we define our set embedding feed-forward network as the maximum-a-posterior (MAP) estimate of the mixture which is approximately attained by a few Expectation-Maximization (EM) steps. The whole MAP-EM steps are differentiable operations with a fixed number of mixture parameters, allowing efficient auto-diff back-propagation for any given downstream task. Furthermore, the proposed mixture set data fitting framework allows unsupervised set representation learning naturally via marginal likelihood maximization aka the empirical Bayes. Interestingly, we also find that OTKE can be seen as a special case of our framework, specifically a single-step EM with extra balanced assignment constraints on the E-step. Compared to OTKE, our approach provides more flexible set embedding as well as prior-induced model regularization. We evaluate our approach on various tasks demonstrating improved performance over the state-of-the-arts.

Overcoming The Spectral Bias of Neural Value Approximation

- Ge Yang, Anurag Ajay, Pulkit Agrawal
- abstract@open-review(Poster): Value approximation using deep neural networks is at the heart of off-policy deep reinforcement learning, and is often the primary module that provides learning signals to the rest of the algorithm. While multi-layer perceptron networks are universal function approximators, recent works in neural kernel regression suggest the presence of a \textit{spectral bias}, where fitting high-frequency components of the value function requires exponentially more gradient update steps than the low-frequency ones. In this work, we re-examine off-policy reinforcement learning through the lens of kernel regression and propose to overcome such bias via a composite neural tangent kernel. With just a single line-change, our approach, the Fourier feature networks (FFN) produce state-of-the-art performance on challenging continuous control domains with only a fraction of the compute.

Faster convergence and better off-policy stability also make it possible to remove the target network without suffering catastrophic divergences, which further reduces TD(0)'s estimation bias on a few tasks. Code and analysis available at https://geyang.github.io/ffn.

Prospect Pruning: Finding Trainable Weights at Initialization using Meta-Gradients

- Milad Alizadeh, Shyam A. Tailor, Luisa M Zintgraf, Joost van Amersfoort, Sebastian Farquhar, Nicholas Donald Lane, Yarin Gal
- abstract@open-review(Poster): Pruning neural networks at initialization would enable us to find sparse models that retain the accuracy of the original network while consuming fewer computational resources for training and inference. However, current methods are insufficient to enable this optimization and lead to a large degradation in model performance. In this paper, we identify a fundamental limitation in the formulation of current methods, namely that their saliency criteria look at a single step at the start of training without taking into account the trainability of the network. While pruning iteratively and gradually has been shown to improve pruning performance, explicit consideration of the training stage that will immediately follow pruning has so far been absent from the computation of the saliency criterion. To overcome the short-sightedness of existing methods, we propose Prospect Pruning (ProsPr), which uses meta-gradients through the first few steps of optimization to determine which weights to prune. ProsPr combines an estimate of the higher-order effects of pruning on the loss and the optimization trajectory to identify the trainable sub-network. Our method achieves state-of-the-art pruning performance on a variety of vision classification tasks, with less data and in a single shot compared to existing pruning-at-initialization methods.

CoMPS: Continual Meta Policy Search

- Glen Berseth, Zhiwei Zhang, Grace Zhang, Chelsea Finn, Sergey Levine
- abstract@open-review(Poster): We develop a new continual meta-learning method to address challenges in sequential multi-task learning. In this setting, the agent's goal is to achieve high reward over any sequence of tasks quickly. Prior meta-reinforcement learning algorithms have demonstrated promising results in accelerating the acquisition of new tasks. However, they require access to all tasks during training. Beyond simply transferring past experience to new tasks, our goal is to devise continual reinforcement learning algorithms that learn to learn, using their experience on previous tasks to learn new tasks more quickly. We introduce a new method, continual meta-policy search (CoMPS), that removes this limitation by meta-training in an incremental fashion, over each task in a sequence, without revisiting prior tasks. CoMPS continuously repeats two subroutines: learning a new task using RL and using the experience from RL to perform completely offline meta-learning to prepare for subsequent task learning. We find that CoMPS outperforms prior continual learning and off-policy meta-reinforcement methods on several sequences of challenging continuous control tasks.

Generalized rectifier wavelet covariance models for texture synthesis

- Antoine Brochard, Sixin Zhang, StÃ©phane Mallat
- abstract@open-review(Poster): State-of-the-art maximum entropy models for texture synthesis are built from statistics relying on image representations defined by convolutional neural networks (CNN). Such representations capture rich structures in texture images, outperforming wavelet-based representations in this regard. However, conversely to neural networks, wavelets offer meaningful representations, as they are known to detect structures at multiple scales (e.g. edges) in images. In this work, we propose a family of statistics built upon non-linear wavelet based representations, that can be viewed as a particular instance of a one-layer CNN, using a generalized rectifier non-linearity. These statistics significantly improve the visual quality of previous classical wavelet-based models, and allow one to produce syntheses of similar quality to state-of-the-art models, on both gray-scale and color textures. We further provide insights on memorization effects in these models.

Towards Evaluating the Robustness of Neural Networks Learned by Transduction

- Jiefeng Chen, Xi Wu, Yang Guo, Yingyu Liang, Somesh Jha
- abstract@open-review(Poster): There has been emerging interest in using transductive learning for adversarial robustness (Goldwasser et al., NeurIPS 2020; Wu et al., ICML 2020; Wang et al., ArXiv 2021). Compared to traditional defenses, these defense mechanisms "dynamically learn" the model based on test-time input; and theoretically, attacking these defenses reduces to solving a bilevel optimization problem, which poses difficulty in crafting adaptive attacks. In this paper, we examine these defense mechanisms from a principled threat analysis perspective. We formulate and analyze threat models for transductive-learning based defenses, and point out important subtleties. We propose the principle of attacking model space for solving bilevel attack objectives, and present Greedy Model Space Attack (GMSA), an attack framework that can serve as a new baseline for evaluating transductive-learning based defenses. Through systematic evaluation, we show that GMSA, even with weak instantiations, can break previous transductive-learning based defenses, which were resilient to previous attacks, such as AutoAttack (Croce and Hein, ICML 2020). On the positive side, we report a somewhat surprising empirical result of "transductive adversarial training": Adversarially retraining the model using fresh randomness at the test time gives a significant increase in robustness against attacks we consider.

OBJECT DYNAMICS DISTILLATION FOR SCENE DECOMPOSITION AND REPRESENTATION

- Qu Tang, Xiangyu Zhu, Zhen Lei, Zhaoxiang Zhang
- abstract@open-review(Poster): The ability to perceive scenes in terms of abstract entities is crucial for us to achieve higher-level intelligence. Recently, several methods have been proposed to learn object-centric representations of scenes with multiple objects, yet most of which focus on static scenes. In this paper, we work on object dynamics and propose Object Dynamics Distillation Network (ODDN), a framework that distillates explicit object dynamics (e.g., velocity) from sequential static representations. ODDN also builds a relation module to model object interactions. We verify our approach on tasks of video reasoning and video prediction, which are two important evaluations for video understanding. The results show that the reasoning model with visual representations of ODDN performs better in answering reasoning questions around physical events in a video compared to the previous state-of-the-art methods. The distilled object dynamics also could be used to predict future video frames given two input frames, involving occlusion and objects collision. In addition, our architecture brings better segmentation quality and higher reconstruction accuracy.

Practical Integration via Separable Bijective Networks

- Christopher M Bender, Patrick Emmanuel, Michael K. Reiter, Junier Oliva
- abstract@open-review(Poster): Neural networks have enabled learning over examples that contain thousands of dimensions. However, most of these models are limited to training and evaluating on a finite collection of \textit{points} and do not consider the hypervolume in which the data resides. Any analysis of the model's local or global behavior is therefore limited to very expensive or imprecise estimators. We propose to formulate neural networks as a composition of a bijective (flow) network followed by a learnable, separable network. This construction allows for learning (or assessing) over full hypervolumes with precise estimators at tractable computational cost via integration over the \textit{input space}. We develop the necessary machinery, propose several practical integrals to use during training, and demonstrate their utility.

Self-Joint Supervised Learning

- Navid Kardan, Mubarak Shah, Mitch Hill
- abstract@open-review(Poster): Supervised learning is a fundamental framework used to train machine learning systems. A supervised learning problem is often formulated using an i.i.d. assumption that restricts model attention to a single relevant signal at a time when predicting. This contrasts with the human ability to actively use related samples as reference when making decisions. We hypothesize that the restriction to a single signal for each prediction

in the standard i.i.d. framework contributes to well-known drawbacks of supervised learning: making overconfident predictions and vulnerability to overfitting, adversarial attacks, and out-of-distribution data. To address these limitations, we propose a new supervised learning paradigm called self-joint learning that generalizes the standard approach by modeling the joint conditional distribution of two observed samples, where each sample is an image and its label. Rather than assuming samples are independent, our models explicitly learn the sample-to-sample relation of conditional independence. Our framework can naturally incorporate auxiliary unlabeled data to further improve the performance. Experiments on benchmark image datasets show our method offers significant improvement over standard supervised learning in terms of accuracy, robustness against adversarial attacks, out-of-distribution detection, and overconfidence mitigation.

Rethinking Supervised Pre-Training for Better Downstream Transferring

- Yutong Feng, Jianwen Jiang, Mingqian Tang, Rong Jin, Yue Gao
- abstract@open-review(Poster): The pretrain-finetune paradigm has shown outstanding performance on many applications of deep learning, where a model is pre-trained on an upstream large dataset (e.g. ImageNet), and is then fine-tuned to different downstream tasks. Though for most cases, the pre-training stage is conducted based on supervised methods, recent works on self-supervised pre-training have shown powerful transferability and even outperform supervised pre-training on multiple downstream tasks. It thus remains an open question how to better generalize supervised pre- training model to downstream tasks. In this paper, we argue that the worse transferability of existing supervised pre-training methods arise from the negligence of valuable intra-class semantic difference. This is because these methods tend to push images from the same class close to each other despite of the large diversity in their visual contents, a problem to which referred as â€œoverfit of upstream tasksâ€ . To alleviate this problem, we propose a new supervised pre-training method based on Leave-One-Out K-Nearest-Neighbor, or LOOK for short. It relieves the problem of overfitting upstream tasks by only requiring each image to share its class label with most of its k nearest neighbors, thus allowing each class to exhibit a multi-mode distribution and consequentially preserving part of intra-class difference for better transferring to downstream tasks. We developed efficient implementation of the proposed method that scales well to large datasets. Experimental studies on multiple downstream tasks show that LOOK outperforms other state-of-the-art methods for supervised and self-supervised pre-training.

A Zest of LIME: Towards Architecture-Independent Model Distances

- Hengrui Jia, Hongyu Chen, Jonas Guan, Ali Shahin Shamsabadi, Nicolas Papernot
- abstract@open-review(Poster): Definitions of the distance between two machine learning models either characterize the similarity of the models' predictions or of their weights. While similarity of weights is attractive because it implies similarity of predictions in the limit, it suffers from being inapplicable to comparing models with different architectures. On the other hand, the similarity of predictions is broadly applicable but depends heavily on the choice of model inputs during comparison. In this paper, we instead propose to compute distance between black-box models by comparing their Local Interpretable Model-Agnostic Explanations (LIME). To compare two models, we take a reference dataset, and locally approximate the models on each reference point with linear models trained by LIME. We then compute the cosine distance between the concatenated weights of the linear models. This yields an approach that is both architecture-independent and possesses the benefits of comparing models in weight space. We empirically show that our method, which we call Zest, can be applied to two problems that require measurements of model similarity: detecting model stealing and machine unlearning.

Meta-Imitation Learning by Watching Video Demonstrations

- Jiayi Li, Tao Lu, Xiaoge Cao, Yinghao Cai, Shuo Wang
- abstract@open-review(Poster): Meta-Imitation Learning is a promising technique for the robot to learn a new task from observing one or a few human demonstrations. However, it usually requires a significant number of demonstrations both from humans and robots during the meta-training phase, which is a laborious and hard work for data collection, especially in recording the actions and specifying the correspondence between human and robot. In this work, we present an approach of meta-imitation learning by watching video demonstrations from humans. In comparison to prior works, our approach is able to translate human videos into practical robot demonstrations and train the meta-policy with adaptive loss based on the quality of the translated data. Our approach relies only on human videos and does not require robot demonstration, which facilitates data collection and is more in line with human imitation behavior. Experiments reveal that our method achieves the comparable performance to the baseline on fast learning a set of vision-based tasks through watching a single video demonstration.

Understanding Intrinsic Robustness Using Label Uncertainty

- Xiao Zhang, David Evans
- abstract@open-review(Poster): A fundamental question in adversarial machine learning is whether a robust classifier exists for a given task. A line of research has made some progress towards this goal by studying the concentration of measure, but we argue standard concentration fails to fully characterize the intrinsic robustness of a classification problem since it ignores data labels which are essential to any classification task. Building on a novel definition of label uncertainty, we empirically demonstrate that error regions induced by state-of-the-art models tend to have much higher label uncertainty than randomly-selected subsets. This observation motivates us to adapt a concentration estimation algorithm to account for label uncertainty, resulting in more accurate intrinsic robustness measures for benchmark image classification problems.

Efficient Split-Mix Federated Learning for On-Demand and In-Situ Customization

- Junyuan Hong, Haotao Wang, Zhangyang Wang, Jiayu Zhou
- abstract@open-review(Poster): Federated learning (FL) provides a distributed learning framework for multiple participants to collaborate learning without sharing raw data. In many practical FL scenarios, participants have heterogeneous resources due to disparities in hardware and inference dynamics that require quickly loading models of different sizes and levels of robustness. The heterogeneity and dynamics together impose significant challenges to existing FL approaches and thus greatly limit FL's applicability. In this paper, we propose a novel Split-Mix FL strategy for heterogeneous participants that, once training is done, provides in-situ customization of model sizes and robustness. Specifically, we achieve customization by learning a set of base sub-networks of different sizes and robustness levels, which are later aggregated on-demand according to inference requirements. This split-mix strategy achieves customization with high efficiency in communication, storage, and inference. Extensive experiments demonstrate that our method provides better in-situ customization than the existing heterogeneous-architecture FL methods. Codes and pre-trained models are available: https://github.com/illidanlab/SplitMix.

Anti-Concentrated Confidence Bonuses For Scalable Exploration

- Jordan T. Ash, Cyril Zhang, Surbhi Goel, Akshay Krishnamurthy, Sham M. Kakade
- abstract@open-review(Poster): Intrinsic rewards play a central role in handling the exploration-exploitation tradeoff when designing sequential decision-making algorithms, in both foundational theory and state-of-the-art deep reinforcement learning. The LinUCB algorithm, a centerpiece of the stochastic linear bandits literature, prescribes an elliptical bonus which addresses the challenge of leveraging shared information in large action spaces. This bonus scheme cannot be directly transferred to high-dimensional exploration problems, however, due to the computational cost of maintaining the inverse covariance matrix of action features. We introduce anti-concentrated confidence bounds for efficiently approximating the elliptical bonus, using an ensemble of regressors trained to predict random noise from policy network-derived features. Using this approximation, we obtain stochastic linear bandit

algorithms which obtain $\tilde O(d \sqrt{T})$ regret bounds for $\mathsf{poly}(d)$ fixed actions. We develop a practical variant that is competitive with contemporary intrinsic reward heuristics on Atari benchmarks.

Sqrt(d) Dimension Dependence of Langevin Monte Carlo

- Ruilin Li, Hongyuan Zha, Molei Tao
- abstract@open-review(Poster): This article considers the popular MCMC method of unadjusted Langevin Monte Carlo (LMC) and provides a non-asymptotic analysis of its sampling error in 2-Wasserstein distance. The proof is based on a refinement of mean-square analysis in Li et al. (2019), and this refined framework automates the analysis of a large class of sampling algorithms based on discretizations of contractive SDEs. Using this framework, we establish an $\tilde{O}(\sqrt{d}/\epsilon)$ mixing time bound for LMC, without warm start, under the common log-smooth and log-strongly-convex conditions, plus a growth condition on the 3rd-order derivative of the potential of target measures. This bound improves the best previously known $\tilde{O}(d/\epsilon)$ result and is optimal (in terms of order) in both dimension d and accuracy tolerance ϵ for target measures satisfying the aforementioned assumptions. Our theoretical analysis is further validated by numerical experiments.

Relational Surrogate Loss Learning

- Tao Huang, Zekang Li, Hua Lu, Yong Shan, Shusheng Yang, Yang Feng, Fei Wang, Shan You, Chang Xu
- abstract@open-review(Poster): Evaluation metrics in machine learning are often hardly taken as loss functions, as they could be non-differentiable and non-decomposable, e.g., average precision and F1 score. This paper aims to address this problem by revisiting the surrogate loss learning, where a deep neural network is employed to approximate the evaluation metrics. Instead of pursuing an exact recovery of the evaluation metric through a deep neural network, we are reminded of the purpose of the existence of these evaluation metrics, which is to distinguish whether one model is better or worse than another. In this paper, we show that directly maintaining the relation of models between surrogate losses and metrics suffices, and propose a rank correlation-based optimization method to maximize this relation and learn surrogate losses. Compared to previous works, our method is much easier to optimize and enjoys significant efficiency and performance gains. Extensive experiments show that our method achieves improvements on various tasks including image classification and neural machine translation, and even outperforms state-of-the-art methods on human pose estimation and machine reading comprehension tasks. Code is available at: https://github.com/hunto/ReLoss.

Structure-Aware Transformer Policy for Inhomogeneous Multi-Task Reinforcement Learning

- Sunghoon Hong, Deunsol Yoon, Kee-Eung Kim
- abstract@open-review(Poster): Modular Reinforcement Learning, where the agent is assumed to be morphologically structured as a graph, for example composed of limbs and joints, aims to learn a policy that is transferable to a structurally similar but different agent. Compared to traditional Multi-Task Reinforcement Learning, this promising approach allows us to cope with inhomogeneous tasks where the state and action space dimensions differ across tasks. Graph Neural Networks are a natural model for representing the pertinent policies, but a recent work has shown that their multi-hop message passing mechanism is not ideal for conveying important information to other modules and thus a transformer model without morphological information was proposed. In this work, we argue that the morphological information is still very useful and propose a transformer policy model that effectively encodes such information. Specifically, we encode the morphological information in terms of the traversal-based positional embedding and the graph-based relational embedding. We empirically show that the morphological information is crucial for modular reinforcement learning, substantially outperforming prior state-of-the-art methods on multi-task learning as well as transfer learning settings with different state and action space dimensions.

Toward Efficient Low-Precision Training: Data Format Optimization and Hysteresis Quantization

- Sunwoo Lee, Jeongwoo Park, Dongsuk Jeon
- abstract@open-review(Poster): As the complexity and size of deep neural networks continue to increase, low-precision training has been extensively studied in the last few years to reduce hardware overhead. Training performance is largely affected by the numeric formats representing different values in low-precision training, but finding an optimal format typically requires numerous training runs, which is a very time-consuming process. In this paper, we propose a method to efficiently find an optimal format for activations and errors without actual training. We employ this method to determine an 8-bit format suitable for training various models. In addition, we propose hysteresis quantization to suppress undesired fluctuation in quantized weights during training. This scheme enables deeply quantized training using 4-bit weights, exhibiting only 0.2% degradation for ResNet-18 trained on ImageNet.

Knowledge Infused Decoding

- Ruibo Liu, Guoqing Zheng, Shashank Gupta, Radhika Gaonkar, Chongyang Gao, Soroush Vosoughi, Milad Shokouhi, Ahmed Hassan Awadallah
- abstract@open-review(Poster): Pre-trained language models (LMs) have been shown to memorize a substantial amount of knowledge from the pre-training corpora; however, they are still limited in recalling factually correct knowledge given a certain context. Hence. they tend to suffer from counterfactual or hallucinatory generation when used in knowledge-intensive natural language generation (NLG) tasks. Recent remedies to this problem focus on modifying either the pre-training or task fine-tuning objectives to incorporate knowledge, which normally require additional costly training or architecture modification of LMs for practical applications.

We present Knowledge Infused Decoding (KID)---a novel decoding algorithm for generative LMs, which dynamically infuses external knowledge into each step of the LM decoding. Specifically, we maintain a local knowledge memory based on the current context, interacting with a dynamically created external knowledge trie, and continuously update the local memory as a knowledge-aware constraint to guide decoding via reinforcement learning. On six diverse knowledge-intensive NLG tasks, task-agnostic LMs (e.g., GPT-2 and BART) armed with KID outperform many task-optimized state-of-the-art models, and show particularly strong performance in few-shot scenarios over seven related knowledge-infusion techniques. Human evaluation confirms KID's ability to generate more relevant and factual language for the input context when compared with multiple baselines. Finally, KID also alleviates exposure bias and provides stable generation quality when generating longer sequences.

Parallel Training of GRU Networks with a Multi-Grid Solver for Long Sequences

- Euhyun Moon, Eric C Cyr
- abstract@open-review(Poster): Parallelizing Gated Recurrent Unit (GRU) is a challenging task, as the training procedure of GRU is inherently sequential. Prior efforts to parallelize GRU have largely focused on conventional parallelization strategies such as data-parallel and model-parallel training algorithms. However, when the given sequences are very long, existing approaches are still inevitably performance limited in terms of both training time and model accuracy. In this paper, we present a novel parallel training scheme (called parallel-in-time) for GRU based on a multigrid reduction in time (MGRIT) solver. MGRIT partitions a sequence into multiple shorter sub-sequences and trains the sub-sequences on different processors in parallel. The key to achieving speedup is a hierarchical correction of the hidden state to accelerate end-to-end communication in both the forward and backward propagation phases of gradient descent. Experimental results on the HMDB51 dataset, where each video is an image sequence, demonstrate that a new parallel training scheme of GRU achieves up to $6.5 \times$ speedup over a serial approach. As efficiency of our new parallelization strategy is associated with the sequence length, our parallel GRU algorithm achieves significant performance improvement as the length of sequence increases. Further, the proposed approach can be applied simultaneously with batch and other forms of model parallelism.

QUERY EFFICIENT DECISION BASED SPARSE ATTACKS AGAINST BLACK-BOX DEEP LEARNING MODELS

- Viet Vo, Ehsan M Abbasnejad, Damith Ranasinghe
- abstract@open-review(Poster): Despite our best efforts, deep learning models remain highly vulnerable to even tiny adversarial perturbations applied to the inputs. The ability to extract information from solely the output of a machine learning model to craft adversarial perturbations to black-box models is a practical threat against real-world systems, such as Machine Learning as a Service (MLaaS), particularly $sparse~attacks$. The realization of sparse attacks in black-box settings demonstrates that machine learning models are more vulnerable than we believe. Because, these attacks aim to $minimize~the~number~of~perturbed~pixels$â€”measured by l_0 normâ€”required to mislead a model by $solely$ observing the decision ($the~predicted~label$) returned to a model query; the so-called $decision-based~setting$. But, such an attack leads to an NP-hard optimization problem. We develop an evolution-based algorithmâ€”$SparseEvo$â€”for the problem and evaluate it against both convolutional deep neural networks and $vision~transformers$. Notably, vision transformers are yet to be investigated under a decision-based attack setting. SparseEvo requires significantly fewer queries than the state-of-the-art sparse attack $Pointwise$ for both untargeted and targeted attacks. The attack algorithm, although conceptually simple, is competitive with only a limited query budget against the state-of-the-art gradient-based $white-box$ attacks in standard computer vision tasks such as $ImageNet$. Importantly, the query efficient SparseEvo, along with decision-based attacks, in general, raises new questions regarding the safety of deployed systems and poses new directions to study and understand the robustness of machine learning models.

Almost Tight L0-norm Certified Robustness of Top-k Predictions against Adversarial Perturbations

- Jinyuan Jia, Binghui Wang, Xiaoyu Cao, Hongbin Liu, Neil Zhenqiang Gong
- abstract@open-review(Poster): Top-k predictions are used in many real-world applications such as machine learning as a service, recommender systems, and web searches. ℓ_0-norm adversarial perturbation characterizes an attack that arbitrarily modifies some features of an input such that a classifier makes an incorrect prediction for the perturbed input. ℓ_0-norm adversarial perturbation is easy to interpret and can be implemented in the physical world. Therefore, certifying robustness of top-k predictions against ℓ_0-norm adversarial perturbation is important. However, existing studies either focused on certifying ℓ_0-norm robustness of top-1 predictions or ℓ_2-norm robustness of top-k predictions. In this work, we aim to bridge the gap. Our approach is based on randomized smoothing, which builds a provably robust classifier from an arbitrary classifier via randomizing an input. Our major theoretical contribution is an almost tight ℓ_0-norm certified robustness guarantee for top-k predictions. We empirically evaluate our method on CIFAR10 and ImageNet. For instance, our method can build a classifier that achieves a certified top-3 accuracy of 69.2\% on ImageNet when an attacker can arbitrarily perturb 5 pixels of a testing image.

Proving the Lottery Ticket Hypothesis for Convolutional Neural Networks

- Arthur da Cunha, Emanuele Natale, Laurent Viennot
- abstract@open-review(Poster): The lottery ticket hypothesis states that a randomly-initialized neural network contains a small subnetwork which, when trained in isolation, can compete with the performance of the original network. Recent theoretical works proved an even stronger version: every sufficiently overparameterized (dense) neural network contains a subnetwork that, even without training, achieves accuracy comparable to that of the trained large network. These works left as an open problem to extend the result to convolutional neural networks (CNNs). In this work we provide such generalization by showing that, with high probability, it is possible to approximate any CNN by pruning a random CNN whose size is larger by a logarithmic factor.

Discovering Nonlinear PDEs from Scarce Data with Physics-encoded Learning

- Chengping Rao, Pu Ren, Yang Liu, Hao Sun
- abstract@open-review(Poster): There have been growing interests in leveraging experimental measurements to discover the underlying partial differential equations (PDEs) that govern complex physical phenomena. Although past research attempts have achieved great success in data-driven PDE discovery, the robustness of the existing methods cannot be guaranteed when dealing with low-quality measurement data. To overcome this challenge, we propose a novel physics-encoded discrete learning framework for discovering spatiotemporal PDEs from scarce and noisy data. The general idea is to (1) firstly introduce a novel deep convolutional-recurrent networks, which can encode prior physics knowledge (e.g., known terms, assumed PDE structure, initial/boundary conditions, etc.) while remaining flexible on representation capability, to accurately reconstruct high-fidelity data, and (2) then perform sparse regression with the reconstructed data to identify the analytical form of the governing PDEs. We validate our proposed framework on three high-dimensional PDE systems. The effectiveness and superiority of the proposed method over baselines are demonstrated.

Rethinking Goal-Conditioned Supervised Learning and Its Connection to Offline RL

- Rui Yang, Yiming Lu, Wenzhe Li, Hao Sun, Meng Fang, Yali Du, Xiu Li, Lei Han, Chongjie Zhang
- abstract@open-review(Poster): Solving goal-conditioned tasks with sparse rewards using self-supervised learning is promising because of its simplicity and stability over current reinforcement learning (RL) algorithms. A recent work, called Goal-Conditioned Supervised Learning (GCSL), provides a new learning framework by iteratively relabeling and imitating self-generated experiences. In this paper, we revisit the theoretical property of GCSL --- optimizing a lower bound of the goal reaching objective, and extend GCSL as a novel offline goal-conditioned RL algorithm. The proposed method is named Weighted GCSL (WGCSL), in which we introduce an advanced compound weight consisting of three parts (1) discounted weight for goal relabeling, (2) goal-conditioned exponential advantage weight, and (3) best-advantage weight. Theoretically, WGCSL is proved to optimize an equivalent lower bound of the goal-conditioned RL objective and generates monotonically improved policies via an iterated scheme. The monotonic property holds for any behavior policies, and therefore WGCSL can be applied to both online and offline settings. To evaluate algorithms in the offline goal-conditioned RL setting, we provide a benchmark including a range of point and simulated robot domains. Experiments in the introduced benchmark demonstrate that WGCSL can consistently outperform GCSL and existing state-of-the-art offline methods in the fully offline goal-conditioned setting.

Topologically Regularized Data Embeddings

- Robin Vandaele, Bo Kang, Jefrey Lijffijt, Tijl De Bie, Yvan Saeys
- abstract@open-review(Poster): Unsupervised feature learning often finds low-dimensional embeddings that capture the structure of complex data. For tasks for which prior expert topological knowledge is available, incorporating this into the learned representation may lead to higher quality embeddings. For example, this may help one to embed the data into a given number of clusters, or to accommodate for noise that prevents one from deriving the distribution of the data over the model directly, which can then be learned more effectively. However, a general tool for integrating different prior topological knowledge into embeddings is lacking. Although differentiable topology layers have been recently developed that can (re)shape embeddings into prespecified topological models, they have two important limitations for representation learning, which we address in this paper. First, the currently suggested topological losses fail to represent simple models such as clusters and flares in a natural manner. Second, these losses neglect all original structural (such as neighborhood) information in the data that is useful for learning. We overcome these limitations by introducing a new set of topological losses, and proposing their usage as a way for topologically regularizing data embeddings to naturally represent a prespecified model. We include thorough experiments on synthetic and real data that highlight the usefulness and versatility of this approach, with applications ranging from modeling high-dimensional single-cell data, to graph embedding.

PF-GNN: Differentiable particle filtering based approximation of universal graph representations

- Mohammed Haroon Dupty, Yanfei Dong, Wee Sun Lee

- abstract@open-review(Poster): Message passing Graph Neural Networks (GNNs) are known to be limited in expressive power by the 1-WL color-refinement test for graph isomorphism. Other more expressive models either are computationally expensive or need preprocessing to extract structural features from the graph. In this work, we propose to make GNNs universal by guiding the learning process with exact isomorphism solver techniques which operate on the paradigm of $\textit{Individualization and refinement}$ (IR), a method to artificially introduce asymmetry and further refine the coloring when 1-WL stops. Isomorphism solvers generate a search-tree of colorings whose leaves uniquely identify the graph. However, the tree grows exponentially large and needs hand-crafted pruning techniques which are not desirable from a learning perspective. We take a probabilistic view and approximate the search tree of colorings (i.e. embeddings) by sampling multiple paths from root to leaves of the search-tree. To learn more discriminative representations, we guide the sampling process with $\textit{particle filter}$ updates, a principled approach for sequential state estimation. Our algorithm is end-to-end differentiable, can be applied with any GNN as backbone and learns richer graph representations with only linear increase in runtime. Experimental evaluation shows that our approach consistently outperforms leading GNN models on both synthetic benchmarks for isomorphism detection as well as real-world datasets.

Nonlinear ICA Using Volume-Preserving Transformations

- Xiaojiang Yang, Yi Wang, Jiacheng Sun, Xing Zhang, Shifeng Zhang, Zhenguo Li, Junchi Yan
- abstract@open-review(Poster): Nonlinear ICA is a fundamental problem in machine learning, aiming to identify the underlying independent components (sources) from data which is assumed to be a nonlinear function (mixing function) of these sources. Recent works prove that if the sources have some particular structures (e.g. temporal structure), they are theoretically identifiable even if the mixing function is arbitrary. However, in many cases such restrictions on the sources are difficult to satisfy or even verify, hence it inhibits the applicability of the proposed methods. Different from these works, we propose a general framework for nonlinear ICA, in which the mixing function is assumed to be a volume-preserving transformation, and meanwhile the conditions on the sources can be much looser. We provide an insightful proof of the identifiability of the proposed framework. We implement the framework by volume-preserving Flow-based models, and verify our theory by experiments on artificial data and synthesized images. Moreover, results on real-world images indicate that our framework can disentangle interpretable features.

Online Ad Hoc Teamwork under Partial Observability

- Pengjie Gu, Mengchen Zhao, Jianye Hao, Bo An
- abstract@open-review(Poster): Autonomous agents often need to work together as a team to accomplish complex cooperative tasks. Due to privacy and other realistic constraints, agents might need to collaborate with previously unknown teammates on the fly. This problem is known as ad hoc teamwork, which remains a core research challenge. Prior works usually rely heavily on strong assumptions like full observability, fixed and predefined teammates' types. This paper relaxes these assumptions with a novel reinforcement learning framework called ODITS, which allows the autonomous agent to adapt to arbitrary teammates in an online fashion. Instead of limiting teammates into a finite set of predefined types, ODITS automatically learns latent variables of teammates' behaviors to infer how to cooperate with new teammates effectively. To overcome partial observability, we introduce an information-based regularizer to derive proxy representations of the learned variables from local observations. Extensive experimental results show that ODITS significantly outperforms various baselines in widely used ad hoc teamwork tasks.

Continual Normalization: Rethinking Batch Normalization for Online Continual Learning

- Quang Pham, Chenghao Liu, Steven HOI
- abstract@open-review(Poster): Existing continual learning methods use Batch Normalization (BN) to facilitate training and improve generalization across tasks. However, the non-i.i.d and non-stationary nature of continual learning data, especially in the online setting, amplify the discrepancy between training and testing in BN and hinder the performance of older tasks. In this work, we study the cross-task normalization effect of BN in online continual learning where BN normalizes the testing data using moments biased towards the current task, resulting in higher catastrophic forgetting. This limitation motivates us to propose a simple yet effective method that we call Continual Normalization (CN) to facilitate training similar to BN while mitigating its negative effect. Extensive experiments on different continual learning algorithms and online scenarios show that CN is a direct replacement for BN and can provide substantial performance improvements. Our implementation will be made publicly available upon acceptance.

Equivariant Graph Mechanics Networks with Constraints

- Wenbing Huang, Jiaqi Han, Yu Rong, Tingyang Xu, Fuchun Sun, Junzhou Huang
- abstract@open-review(Poster): Learning to reason about relations and dynamics over multiple interacting objects is a challenging topic in machine learning. The challenges mainly stem from that the interacting systems are exponentially-compositional, symmetrical, and commonly geometrically-constrained. Current methods, particularly the ones based on equivariant Graph Neural Networks (GNNs), have targeted on the first two challenges but remain immature for constrained systems. In this paper, we propose Graph Mechanics Network (GMN) which is combinatorially efficient, equivariant and constraint-aware. The core of GMN is that it represents, by generalized coordinates, the forward kinematics information (positions and velocities) of a structural object. In this manner, the geometrical constraints are implicitly and naturally encoded in the forward kinematics. Moreover, to allow equivariant message passing in GMN, we have developed a general form of orthogonality-equivariant functions, given that the dynamics of constrained systems are more complicated than the unconstrained counterparts. Theoretically, the proposed equivariant formulation is proved to be universally expressive under certain conditions. Extensive experiments support the advantages of GMN compared to the state-of-the-art GNNs in terms of prediction accuracy, constraint satisfaction and data efficiency on the simulated systems consisting of particles, sticks and hinges, as well as two real-world datasets for molecular dynamics prediction and human motion capture.

Towards Continual Knowledge Learning of Language Models

- Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun KIM, Stanley Jungkyu Choi, Minjoon Seo
- abstract@open-review(Poster): Large Language Models (LMs) are known to encode world knowledge in their parameters as they pretrain on a vast amount of web corpus, which is often utilized for performing knowledge-dependent downstream tasks such as question answering, fact-checking, and open dialogue. In real-world scenarios, the world knowledge stored in the LMs can quickly become outdated as the world changes, but it is non-trivial to avoid catastrophic forgetting and reliably acquire new knowledge while preserving invariant knowledge. To push the community towards better maintenance of ever-changing LMs, we formulate a new continual learning (CL) problem called Continual Knowledge Learning (CKL). We construct a new benchmark and metric to quantify the retention of time-invariant world knowledge, the update of outdated knowledge, and the acquisition of new knowledge. We adopt applicable recent methods from literature to create several strong baselines. Through extensive experiments, we find that CKL exhibits unique challenges that are not addressed in previous CL setups, where parameter expansion is necessary to reliably retain and learn knowledge simultaneously. By highlighting the critical causes of knowledge forgetting, we show that CKL is a challenging and important problem that helps us better understand and train ever-changing LMs.

Surreal-GAN:Semi-Supervised Representation Learning via GAN for uncovering heterogeneous disease-related imaging patterns

- Zhijian Yang, Junhao Wen, Christos Davatzikos
- abstract@open-review(Poster): A plethora of machine learning methods have been applied to imaging data, enabling the construction of clinically relevant imaging signatures of neurological and neuropsychiatric diseases. Oftentimes, such methods don't explicitly model the heterogeneity of disease effects, or approach it via nonlinear models that are not interpretable. Moreover, unsupervised methods may parse heterogeneity that is driven by nuisance

confounding factors that affect brain structure or function, rather than heterogeneity relevant to a pathology of interest. On the other hand, semi-supervised clustering methods seek to derive a dichotomous subtype membership, ignoring the truth that disease heterogeneity spatially and temporally extends along a continuum. To address the aforementioned limitations, herein, we propose a novel method, termed Surreal-GAN (Semi-SUpeRvised ReprEsentAtion Learning via GAN). Using cross-sectional imaging data, Surreal-GAN dissects underlying disease-related heterogeneity under the principle of semi-supervised clustering (cluster mappings from normal control to patient), proposes a continuously dimensional representation, and infers the disease severity of patients at individual level along each dimension. The model first learns a transformation function from normal control (CN) domain to the patient (PT) domain with latent variables controlling transformation directions. An inverse mapping function together with regularization on function continuity, pattern orthogonality and monotonicity was also imposed to make sure that the transformation function captures necessarily meaningful imaging patterns with clinical significance. We first validated the model through extensive semi-synthetic experiments, and then demonstrate its potential in capturing biologically plausible imaging patterns in Alzheimer's disease (AD).

SURF: Semi-supervised Reward Learning with Data Augmentation for Feedback-efficient Preference-based Reinforcement Learning

- Jongjin Park, Younggyo Seo, Jinwoo Shin, Honglak Lee, Pieter Abbeel, Kimin Lee
- abstract@open-review(Poster): Preference-based reinforcement learning (RL) has shown potential for teaching agents to perform the target tasks without a costly, pre-defined reward function by learning the reward with a supervisorâ€™s preference between the two agent behaviors. However, preference-based learning often requires a large amount of human feedback, making it difficult to apply this approach to various applications. This data-efficiency problem, on the other hand, has been typically addressed by using unlabeled samples or data augmentation techniques in the context of supervised learning. Motivated by the recent success of these approaches, we present SURF, a semi-supervised reward learning framework that utilizes a large amount of unlabeled samples with data augmentation. In order to leverage unlabeled samples for reward learning, we infer pseudo-labels of the unlabeled samples based on the confidence of the preference predictor. To further improve the label-efficiency of reward learning, we introduce a new data augmentation that temporally crops consecutive subsequences from the original behaviors. Our experiments demonstrate that our approach significantly improves the feedback-efficiency of the state-of-the-art preference-based method on a variety of locomotion and robotic manipulation tasks.

Convergent Graph Solvers

- Junyoung Park, Jinhyun Choo, Jinkyoo Park
- abstract@open-review(Poster): We propose the convergent graph solver (CGS), a deep learning method that learns iterative mappings to predict the properties of a graph system at its stationary state (fixed point) with guaranteed convergence. The forward propagation of CGS proceeds in three steps: (1) constructing the input-dependent linear contracting iterative maps, (2) computing the fixed points of the iterative maps, and (3) decoding the fixed points to estimate the properties. The contractivity of the constructed linear maps guarantees the existence and uniqueness of the fixed points following the Banach fixed point theorem. To train CGS efficiently, we also derive a tractable analytical expression for its gradient by leveraging the implicit function theorem. We evaluate the performance of CGS by applying it to various network-analytic and graph benchmark problems. The results indicate that CGS has competitive capabilities for predicting the stationary properties of graph systems, irrespective of whether the target systems are linear or non-linear. CGS also shows high performance for graph classification problems where the existence or the meaning of a fixed point is hard to be clearly defined, which highlights the potential of CGS as a general graph neural network architecture.

Spread Spurious Attribute: Improving Worst-group Accuracy with Spurious Attribute Estimation

- Junhyun Nam, Jaehyung Kim, Jaeho Lee, Jinwoo Shin
- abstract@open-review(Poster): The paradigm of worst-group loss minimization has shown its promise in avoiding to learn spurious correlations, but requires costly additional supervision on spurious attributes. To resolve this, recent works focus on developing weaker forms of supervision---e.g., hyperparameters discovered with a small number of validation samples with spurious attribute annotation---but none of the methods retain comparable performance to methods using full supervision on the spurious attribute. In this paper, instead of searching for weaker supervisions, we ask: Given access to a fixed number of samples with spurious attribute annotations, what is the best achievable worst-group loss if we "fully exploit" them? To this end, we propose a pseudo-attribute-based algorithm, coined Spread Spurious Attribute (SSA), for improving the worst-group accuracy. In particular, we leverage samples both with and without spurious attribute annotations to train a model to predict the spurious attribute, then use the pseudo-attribute predicted by the trained model as supervision on the spurious attribute to train a new robust model having minimal worst-group loss. Our experiments on various benchmark datasets show that our algorithm consistently outperforms the baseline methods using the same number of validation samples with spurious attribute annotations. We also demonstrate that the proposed SSA can achieve comparable performances to methods using full (100%) spurious attribute supervision, by using a much smaller number of annotated samples---from 0.6% and up to 1.5%, depending on the dataset.

Learning Scenario Representation for Solving Two-stage Stochastic Integer Programs

- Yaoxin Wu, Wen Song, Zhiguang Cao, Jie Zhang
- abstract@open-review(Poster): Many practical combinatorial optimization problems under uncertainty can be modeled as stochastic integer programs (SIPs), which are extremely challenging to solve due to the high complexity. To solve two-stage SIPs efficiently, we propose a conditional variational autoencoder (CVAE) based method to learn scenario representation for a class of SIP instances. Specifically, we design a graph convolutional network based encoder to embed each scenario with the deterministic part of its instance (i.e. context) into a low-dimensional latent space, from which a decoder reconstructs the scenario from its latent representation conditioned on the context. Such a design effectively captures the dependencies of the scenarios on their corresponding instances. We apply the trained encoder to two tasks in typical SIP solving, i.e. scenario reduction and objective prediction. Experiments on two SIP problems show that the learned latent representation significantly boosts the solving performance to attain high-quality solutions in short computational time, and generalizes fairly well to problems of larger sizes or with more scenarios.

Generalization Through the Lens of Leave-One-Out Error

- Gregor Bachmann, Thomas Hofmann, Aurelien Lucchi
- abstract@open-review(Poster): Despite the tremendous empirical success of deep learning models to solve various learning tasks, our theoretical understanding of their generalization ability is very limited. Classical generalization bounds based on tools such as the VC dimension or Rademacher complexity, are so far unsuitable for deep models and it is doubtful that these techniques can yield tight bounds even in the most idealistic settings~\citep{nagarajan2019uniform}. In this work, we instead revisit the concept of leave-one-out (LOO) error to measure the generalization ability of deep models in the so-called kernel regime. While popular in statistics, the LOO error has been largely overlooked in the context of deep learning. By building upon the recently established connection between neural networks and kernel learning, we leverage the closed-form expression for the leave-one-out error, giving us access to an efficient proxy for the test error. We show both theoretically and empirically that the leave-one-out error is capable of capturing various phenomena in generalization theory, such as double descent, random labels or transfer learning. Our work therefore demonstrates that the leave-one-out error provides a tractable way to estimate the generalization ability of deep neural networks in the kernel regime, opening the door to potential, new research directions in the field of generalization.

Self-Supervised Inference in State-Space Models

- David Ruhe, Patrick ForrÃ©

- abstract@open-review(Poster): We perform approximate inference in state-space models with nonlinear state transitions. Without parameterizing a generative model, we apply Bayesian update formulas using a local linearity approximation parameterized by neural networks. It comes accompanied by a maximum likelihood objective that requires no supervision via uncorrupt observations or ground truth latent states. The optimization backpropagates through a recursion similar to the classical Kalman filter and smoother. Additionally, using an approximate conditional independence, we can perform smoothing without having to parameterize a separate model. In scientific applications, domain knowledge can give a linear approximation of the latent transition maps, which we can easily incorporate into our model. Usage of such domain knowledge is reflected in excellent results (despite our model's simplicity) on the chaotic Lorenz system compared to fully supervised and variational inference methods. Finally, we show competitive results on an audio denoising experiment.

On the Role of Neural Collapse in Transfer Learning

- Tomer Galanti, AndrÃ¡s GyÃ¶rgy, Marcus Hutter
- abstract@open-review(Poster): We study the ability of foundation models to learn representations for classification that are transferable to new, unseen classes. Recent results in the literature show that representations learned by a single classifier over many classes are competitive on few-shot learning problems with representations learned by special-purpose algorithms designed for such problems. In this paper, we provide an explanation for this behavior based on the recently observed phenomenon that the features learned by overparameterized classification networks show an interesting clustering property, called neural collapse. We demonstrate both theoretically and empirically that neural collapse generalizes to new samples from the training classes, and -- more importantly -- to new classes as well, allowing foundation models to provide feature maps that work well in transfer learning and, specifically, in the few-shot setting.

Information-theoretic Online Memory Selection for Continual Learning

- Shengyang Sun, Daniele Calandriello, Huiyi Hu, Ang Li, Michalis Titsias
- abstract@open-review(Poster): A challenging problem in task-free continual learning is the online selection of a representative replay memory from data streams. In this work, we investigate the online memory selection problem from an information-theoretic perspective. To gather the most information, we propose the \textit{surprise} and the \textit{learnability} criteria to pick informative points and to avoid outliers. We present a Bayesian model to compute the criteria efficiently by exploiting rank-one matrix structures. We demonstrate that these criteria encourage selecting informative points in a greedy algorithm for online memory selection. Furthermore, by identifying the importance of \textit{the timing to update the memory}, we introduce a stochastic information-theoretic reservoir sampler (InfoRS), which conducts sampling among selective points with high information. Compared to reservoir sampling, InfoRS demonstrates improved robustness against data imbalance. Finally, empirical performances over continual learning benchmarks manifest its efficiency and efficacy.

Dealing with Non-Stationarity in MARL via Trust-Region Decomposition

- Wenhao Li, Xiangfeng Wang, Bo Jin, Junjie Sheng, Hongyuan Zha
- abstract@open-review(Poster): Non-stationarity is one thorny issue in cooperative multi-agent reinforcement learning (MARL). One of the reasons is the policy changes of agents during the learning process. Some existing works have discussed various consequences caused by non-stationarity with several kinds of measurement indicators. This makes the objectives or goals of existing algorithms are inevitably inconsistent and disparate. In this paper, we introduce a novel notion, the δ-$stationarity$ measurement, to explicitly measure the non-stationarity of a policy sequence, which can be further proved to be bounded by the KL-divergence of consecutive joint policies. A straightforward but highly non-trivial way is to control the joint policies' divergence, which is difficult to estimate accurately by imposing the trust-region constraint on the joint policy. Although it has lower computational complexity to decompose the joint policy and impose trust-region constraints on the factorized policies, simple policy factorization like mean-field approximation will lead to more considerable policy divergence, which can be considered as the trust-region decomposition dilemma. We model the joint policy as a pairwise Markov random field and propose a trust-region decomposition network (TRD-Net) based on message passing to estimate the joint policy divergence more accurately. The Multi-Agent Mirror descent policy algorithm with Trust region decomposition, called MAMT, is established by adjusting the trust-region of the local policies adaptively in an end-to-end manner. MAMT can approximately constrain the consecutive joint policies' divergence to satisfy δ-stationarity and alleviate the non-stationarity problem. Our method can bring noticeable and stable performance improvement compared with baselines in cooperative tasks of different complexity.

Information Bottleneck: Exact Analysis of (Quantized) Neural Networks

- Stephan Sloth Lorenzen, Christian Igel, Mads Nielsen
- abstract@open-review(Poster): The information bottleneck (IB) principle has been suggested as a way to analyze deep neural networks. The learning dynamics are studied by inspecting the mutual information (MI) between the hidden layers and the input and output. Notably, separate fitting and compression phases during training have been reported. This led to some controversy including claims that the observations are not reproducible and strongly dependent on the type of activation function used as well as on the way the MI is estimated. Our study confirms that different ways of binning when computing the MI lead to qualitatively different results, either supporting or refusing IB conjectures. To resolve the controversy, we study the IB principle in settings where MI is non-trivial and can be computed exactly. We monitor the dynamics of quantized neural networks, that is, we discretize the whole deep learning system so that no approximation is required when computing the MI. This allows us to quantify the information flow without measurement errors. In this setting, we observed a fitting phase for all layers and a compression phase for the output layer in all experiments; the compression in the hidden layers was dependent on the type of activation function. Our study shows that the initial IB results were not artifacts of binning when computing the MI. However, the critical claim that the compression phase may not be observed for some networks also holds true.

GLASS: GNN with Labeling Tricks for Subgraph Representation Learning

- Xiyuan Wang, Muhan Zhang
- abstract@open-review(Poster): Despite the remarkable achievements of Graph Neural Networks (GNNs) on graph representation learning, few works have tried to use them to predict properties of subgraphs in the whole graph. The existing state-of-the-art method SubGNN introduces an overly complicated subgraph-level GNN model which synthesizes three artificial channels each of which has two carefully designed subgraph-level message passing modules, yet only slightly outperforms a plain GNN which performs node-level message passing and then pools node embeddings within the subgraph. By analyzing SubGNN and plain GNNs, we find that the key for subgraph representation learning might be to distinguish nodes inside and outside the subgraph. With this insight, we propose an expressive and scalable labeling trick, namely max-zero-one, to enhance plain GNNs for subgraph tasks. The resulting model is called GLASS (GNN with LAbeling trickS for Subgraph). We theoretically characterize GLASS's expressive power. Compared with SubGNN, GLASS is more expressive, more scalable, and easier to implement. Experiments on eight benchmark datasets show that GLASS outperforms the strongest baseline by 14.8% on average. And ablation analysis shows that our max-zero-one labeling trick can boost the performance of a plain GNN by up to 105% in maximum, which illustrates the effectiveness of labeling trick on subgraph tasks. Furthermore, training a GLASS model only takes 37% time needed for a SubGNN on average.

MoReL: Multi-omics Relational Learning

- Arman Hasanzadeh, Ehsan Hajiramezanali, Nick Duffield, Xiaoning Qian

- abstract@open-review(Poster): Multi-omics data analysis has the potential to discover hidden molecular interactions, revealing potential regulatory and/or signal transduction pathways for cellular processes of interest when studying life and disease systems. One of critical challenges when dealing with real-world multi-omics data is that they may manifest heterogeneous structures and data quality as often existing data may be collected from different subjects under different conditions for each type of omics data. We propose a novel deep Bayesian generative model to efficiently infer a multi-partite graph encoding molecular interactions across such heterogeneous views, using a fused Gromov-Wasserstein (FGW) regularization between latent representations of corresponding views for integrative analysis. With such an optimal transport regularization in the deep Bayesian generative model, it not only allows incorporating view-specific side information, either with graph-structured or unstructured data in different views, but also increases the model flexibility with the distribution-based regularization. This allows efficient alignment of heterogeneous latent variable distributions to derive reliable interaction predictions compared to the existing point-based graph embedding methods. Our experiments on several real-world datasets demonstrate enhanced performance of MoReL in inferring meaningful interactions compared to existing baselines.

Provable Learning-based Algorithm For Sparse Recovery

- Xinshi Chen, Haoran Sun, Le Song
- abstract@open-review(Poster): Recovering sparse parameters from observational data is a fundamental problem in machine learning with wide applications. Many classic algorithms can solve this problem with theoretical guarantees, but their performances rely on choosing the correct hyperparameters. Besides, hand-designed algorithms do not fully exploit the particular problem distribution of interest. In this work, we propose a deep learning method for algorithm learning called PLISA (Provable Learning-based Iterative Sparse recovery Algorithm). PLISA is designed by unrolling a classic path-following algorithm for sparse recovery, with some components being more flexible and learnable. We theoretically show the improved recovery accuracy achievable by PLISA. Furthermore, we analyze the empirical Rademacher complexity of PLISA to characterize its generalization ability to solve new problems outside the training set. This paper contains novel theoretical contributions to the area of learning-based algorithms in the sense that (i) PLISA is generically applicable to a broad class of sparse estimation problems, (ii) generalization analysis has received less attention so far, and (iii) our analysis makes novel connections between the generalization ability and algorithmic properties such as stability and convergence of the unrolled algorithm, which leads to a tighter bound that can explain the empirical observations. The techniques could potentially be applied to analyze other learning-based algorithms in the literature.

Defending Against Image Corruptions Through Adversarial Augmentations

- Dan Andrei Calian, Florian Stimberg, Olivia Wiles, Sylvestre-Alvise Rebuffi, AndrÃ¡s GyÃ¶rgy, Timothy A Mann, Sven Gowal
- abstract@open-review(Poster): Modern neural networks excel at image classification, yet they remain vulnerable to common image corruptions such as blur, speckle noise or fog. Recent methods that focus on this problem, such as AugMix and DeepAugment, introduce defenses that operate in expectation over a distribution of image corruptions. In contrast, the literature on Lp-norm bounded perturbations focuses on defenses against worst-case corruptions. In this work, we reconcile both approaches by proposing AdversarialAugment, a technique which optimizes the parameters of image-to-image models to generate adversarially corrupted augmented images. We theoretically motivate our method and give sufficient conditions for the consistency of its idealized version as well as that of DeepAugment. Our classifiers improve upon the state-of-the-art on common image corruption benchmarks conducted in expectation on CIFAR-10-C and improve worst-case performance against Lp-norm bounded perturbations on both CIFAR-10 and ImageNet.

Attacking deep networks with surrogate-based adversarial black-box methods is easy

- Nicholas A. Lord, Romain Mueller, Luca Bertinetto
- abstract@open-review(Poster): A recent line of work on black-box adversarial attacks has revived the use of transfer from surrogate models by integrating it into query-based search. However, we find that existing approaches of this type underperform their potential, and can be overly complicated besides. Here, we provide a short and simple algorithm which achieves state-of-the-art results through a search which uses the surrogate network's class-score gradients, with no need for other priors or heuristics. The guiding assumption of the algorithm is that the studied networks are in a fundamental sense learning similar functions, and that a transfer attack from one to the other should thus be fairly "easy". This assumption is validated by the extremely low query counts and failure rates achieved: e.g. an untargeted attack on a VGG-16 ImageNet network using a ResNet-152 as the surrogate yields a median query count of 6 at a success rate of 99.9%. Code is available at https://github.com/fiveai/GFCS.

Autoregressive Diffusion Models

- Emiel Hoogeboom, Alexey A. Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, Tim Salimans
- abstract@open-review(Poster): We introduce Autoregressive Diffusion Models (ARDMs), a model class encompassing and generalizing order-agnostic autoregressive models (Uria et al., 2014) and absorbing discrete diffusion (Austin et al., 2021), which we show are special cases of ARDMs under mild assumptions. ARDMs are simple to implement and easy to train. Unlike standard ARMs, they do not require causal masking of model representations, and can be trained using an efficient objective similar to modern probabilistic diffusion models that scales favourably to highly-dimensional data. At test time, ARDMs support parallel generation which can be adapted to fit any given generation budget. We find that ARDMs require significantly fewer steps than discrete diffusion models to attain the same performance. Finally, we apply ARDMs to lossless compression, and show that they are uniquely suited to this task. Contrary to existing approaches based on bits-back coding, ARDMs obtain compelling results not only on complete datasets, but also on compressing single data points. Moreover, this can be done using a modest number of network calls for (de)compression due to the model's adaptable parallel generation.

Auto-scaling Vision Transformers without Training

- Wuyang Chen, Wei Huang, Xianzhi Du, Xiaodan Song, Zhangyang Wang, Denny Zhou
- abstract@open-review(Poster): This work targets automated designing and scaling of Vision Transformers (ViTs). The motivation comes from two pain spots: 1) the lack of efficient and principled methods for designing and scaling ViTs; 2) the tremendous computational cost of training ViT that is much heavier than its convolution counterpart. To tackle these issues, we propose As-ViT, an auto-scaling framework for ViTs without training, which automatically discovers and scales up ViTs in an efficient and principled manner. Specifically, we first design a "seed" ViT topology by leveraging a training-free search process. This extremely fast search is fulfilled by a comprehensive study of ViT's network complexity, yielding a strong Kendall-tau correlation with ground-truth accuracies. Second, starting from the "seed" topology, we automate the scaling rule for ViTs by growing widths/depths to different ViT layers. This results in a series of architectures with different numbers of parameters in a single run. Finally, based on the observation that ViTs can tolerate coarse tokenization in early training stages, we propose a progressive tokenization strategy to train ViTs faster and cheaper. As a unified framework, As-ViT achieves strong performance on classification (83.5% top1 on ImageNet-1k) and detection (52.7% mAP on COCO) without any manual crafting nor scaling of ViT architectures: the end-to-end model design and scaling process costs only 12 hours on one V100 GPU. Our code is available at https://github.com/VITA-Group/AsViT.

Fine-grained Differentiable Physics: A Yarn-level Model for Fabrics

- Deshan Gong, Zhanxing Zhu, Andrew J. Bulpitt, He Wang
- abstract@open-review(Poster): Differentiable physics modeling combines physics models with gradient-based learning to provide model explicability and data efficiency. It has been used to learn dynamics, solve inverse problems and facilitate design, and is at its inception of impact. Current successes have concentrated on general physics models such as rigid bodies, deformable sheets, etc, assuming relatively simple structures and forces. Their granularity is

intrinsically coarse and therefore incapable of modelling complex physical phenomena. Fine-grained models are still to be developed to incorporate sophisticated material structures and force interactions with gradient-based learning. Following this motivation, we propose a new differentiable fabrics model for composite materials such as cloths, where we dive into the granularity of yarns and model individual yarn physics and yarn-to-yarn interactions. To this end, we propose several differentiable forces, whose counterparts in empirical physics are indifferentiable, to facilitate gradient-based learning. These forces, albeit applied to cloths, are ubiquitous in various physical systems. Through comprehensive evaluation and comparison, we demonstrate our model's $\textit{explicability}$ in learning meaningful physical parameters, $\textit{versatility}$ in incorporating complex physical structures and heterogeneous materials, $\textit{data-efficiency}$ in learning, and $\textit{high-fidelity}$ in capturing subtle dynamics.

Revisiting flow generative models for Out-of-distribution detection

- Dihong Jiang, Sun Sun, Yaoliang Yu
- abstract@open-review(Poster): Deep generative models have been widely used in practical applications such as the detection of out-of-distribution (OOD) data. In this work, we aim to re-examine the potential of generative flow models in OOD detection. We first propose a simple combination of univariate one-sample statistical test (e.g., Kolmogorov-Smirnov) and random projections in the latent space of flow models to perform OOD detection. Then, we propose a two-sample version of our test to account for imperfect flow models. Quite distinctly, our method does not pose parametric assumptions on OOD data and is capable of exploiting any flow model. Experimentally, firstly we confirm the efficacy of our method against state-of-the-art baselines through extensive experiments on several image datasets; secondly we investigate the relationship between model accuracy (e.g., the generation quality) and the OOD detection performance, and found surprisingly that they are not always positively correlated; and thirdly we show that detection in the latent space of flow models generally outperforms detection in the sample space across various OOD datasets, hence highlighting the benefits of training a flow model.

Missingness Bias in Model Debugging

- Saachi Jain, Hadi Salman, Eric Wong, Pengchuan Zhang, Vibhav Vineet, Sai Vemprala, Aleksander Madry
- abstract@open-review(Poster): Missingness, or the absence of features from an input, is a concept fundamental to many model debugging tools. However, in computer vision, pixels cannot simply be removed from an image. One thus tends to resort to heuristics such as blacking out pixels, which may in turn introduce bias into the debugging process. We study such biases and, in particular, show how transformer-based architectures can enable a more natural implementation of missingness, which side-steps these issues and improves the reliability of model debugging in practice.

Meta Learning Low Rank Covariance Factors for Energy Based Deterministic Uncertainty

- Jeffrey Ryan Willette, Hae Beom Lee, Juho Lee, Sung Ju Hwang
- abstract@open-review(Poster): Numerous recent works utilize bi-Lipschitz regularization of neural network layers to preserve relative distances between data instances in the feature spaces of each layer. This distance sensitivity with respect to the data aids in tasks such as uncertainty calibration and out-of-distribution (OOD) detection. In previous works, features extracted with a distance sensitive model are used to construct feature covariance matrices which are used in deterministic uncertainty estimation or OOD detection. However, in cases where there is a distribution over tasks, these methods result in covariances which are sub-optimal, as they may not leverage all of the meta information which can be shared among tasks. With the use of an attentive set encoder, we propose to meta learn either diagonal or diagonal plus low-rank factors to efficiently construct task specific covariance matrices. Additionally, we propose an inference procedure which utilizes scaled energy to achieve a final predictive distribution which is well calibrated under a distributional dataset shift.

Conditional Object-Centric Learning from Video

- Thomas Kipf, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, Klaus Greff
- abstract@open-review(Poster): Object-centric representations are a promising path toward more systematic generalization by providing flexible abstractions upon which compositional world models can be built. Recent work on simple 2D and 3D datasets has shown that models with object-centric inductive biases can learn to segment and represent meaningful objects from the statistical structure of the data alone without the need for any supervision. However, such fully-unsupervised methods still fail to scale to diverse realistic data, despite the use of increasingly complex inductive biases such as priors for the size of objects or the 3D geometry of the scene. In this paper, we instead take a weakly-supervised approach and focus on how 1) using the temporal dynamics of video data in the form of optical flow and 2) conditioning the model on simple object location cues can be used to enable segmenting and tracking objects in significantly more realistic synthetic data. We introduce a sequential extension to Slot Attention which we train to predict optical flow for realistic looking synthetic scenes and show that conditioning the initial state of this model on a small set of hints, such as center of mass of objects in the first frame, is sufficient to significantly improve instance segmentation. These benefits generalize beyond the training distribution to novel objects, novel backgrounds, and to longer video sequences. We also find that such initial-state-conditioning can be used during inference as a flexible interface to query the model for specific objects or parts of objects, which could pave the way for a range of weakly-supervised approaches and allow more effective interaction with trained models.

Scale Efficiently: Insights from Pretraining and Finetuning Transformers

- Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, Donald Metzler
- abstract@open-review(Poster): There remain many open questions pertaining to the scaling behaviour of Transformer architectures. These scaling decisions and findings can be critical, as training runs often come with an associated computational cost which have both financial and/or environmental impact. The goal of this paper is to present scaling insights from pretraining and finetuning Transformers. While Kaplan et al. presents a comprehensive study of the scaling behaviour of Transformer language models, the scope is only on the upstream (pretraining) loss. Therefore, it is still unclear if these set of findings transfer to downstream task within the context of the pretrain-finetune paradigm. The key findings of this paper are as follows: (1) we show that aside from only the model size, model shape matters for downstream fine-tuning, (2) scaling protocols operate differently at different compute regions, (3) widely adopted T5-base and T5-large sizes are Pareto-inefficient. To this end, we present improved scaling protocols whereby our redesigned models achieve similar downstream fine-tuning quality while having 50\% fewer parameters and training 40\% faster compared to the widely adopted T5-base model. We publicly release over 100 pretrained checkpoints of different T5 configurations to facilitate future research and analysis.

Vitruvion: A Generative Model of Parametric CAD Sketches

- Ari Seff, Wenda Zhou, Nick Richardson, Ryan P Adams
- abstract@open-review(Poster): Parametric computer-aided design (CAD) tools are the predominant way that engineers specify physical structures, from bicycle pedals to airplanes to printed circuit boards. The key characteristic of parametric CAD is that design intent is encoded not only via geometric primitives, but also by parameterized constraints between the elements. This relational specification can be viewed as the construction of a constraint program, allowing edits to coherently propagate to other parts of the design. Machine learning offers the intriguing possibility of accelerating the design process via generative modeling of these structures, enabling new tools such as autocompletion, constraint inference, and conditional synthesis. In this work, we present such an approach to generative modeling of parametric CAD sketches, which constitute the basic computational building blocks of modern mechanical design. Our model, trained on real-world designs from the SketchGraphs dataset, autoregressively synthesizes sketches as sequences

of primitives, with initial coordinates, and constraints that reference back to the sampled primitives. As samples from the model match the constraint graph representation used in standard CAD software, they may be directly imported, solved, and edited according to downstream design tasks. In addition, we condition the model on various contexts, including partial sketches (primers) and images of hand-drawn sketches. Evaluation of the proposed approach demonstrates its ability to synthesize realistic CAD sketches and its potential to aid the mechanical design workflow.

Space-Time Graph Neural Networks

- Samar Hadou, Charilaos I Kanatsoulis, Alejandro Ribeiro
- abstract@open-review(Poster): We introduce space-time graph neural network (ST-GNN), a novel GNN architecture, tailored to jointly process the underlying space-time topology of time-varying network data. The cornerstone of our proposed architecture is the composition of time and graph convolutional filters followed by pointwise nonlinear activation functions. We introduce a generic definition of convolution operators that mimic the diffusion process of signals over its underlying support. On top of this definition, we propose space-time graph convolutions that are built upon a composition of time and graph shift operators. We prove that ST-GNNs with multivariate integral Lipschitz filters are stable to small perturbations in the underlying graphs as well as small perturbations in the time domain caused by time warping. Our analysis shows that small variations in the network topology and time evolution of a system does not significantly affect the performance of ST-GNNs. Numerical experiments with decentralized control systems showcase the effectiveness and stability of the proposed ST-GNNs.

Scattering Networks on the Sphere for Scalable and Rotationally Equivariant Spherical CNNs

- Jason McEwen, Christopher Wallis, Augustine N. Mavor-Parker
- abstract@open-review(Poster): Convolutional neural networks (CNNs) constructed natively on the sphere have been developed recently and shown to be highly effective for the analysis of spherical data. While an efficient framework has been formulated, spherical CNNs are nevertheless highly computationally demanding; typically they cannot scale beyond spherical signals of thousands of pixels. We develop scattering networks constructed natively on the sphere that provide a powerful representational space for spherical data. Spherical scattering networks are computationally scalable and exhibit rotational equivariance, while their representational space is invariant to isometries and provides efficient and stable signal representations. By integrating scattering networks as an additional type of layer in the generalized spherical CNN framework, we show how they can be leveraged to scale spherical CNNs to the high-resolution data typical of many practical applications, with spherical signals of many tens of megapixels and beyond.

Bayesian Neural Network Priors Revisited

- Vincent Fortuin, AdriÃ Garriga-Alonso, Sebastian W. Ober, Florian Wenzel, Gunnar Ratsch, Richard E Turner, Mark van der Wilk, Laurence Aitchison
- abstract@open-review(Poster): Isotropic Gaussian priors are the de facto standard for modern Bayesian neural network inference. However, it is unclear whether these priors accurately reflect our true beliefs about the weight distributions or give optimal performance. To find better priors, we study summary statistics of neural network weights in networks trained using stochastic gradient descent (SGD). We find that convolutional neural network (CNN) and ResNet weights display strong spatial correlations, while fully connected networks (FCNNs) display heavy-tailed weight distributions. We show that building these observations into priors can lead to improved performance on a variety of image classification datasets. Surprisingly, these priors mitigate the cold posterior effect in FCNNs, but slightly increase the cold posterior effect in ResNets.

Goal-Directed Planning via Hindsight Experience Replay

- Lorenzo Moro, Amarildo Likmeta, Enrico Prati, Marcello Restelli
- abstract@open-review(Poster): We consider the problem of goal-directed planning under a deterministic transition model. Monte Carlo Tree Search has shown remarkable performance in solving deterministic control problems. It has been extended from complex continuous domains through function approximators to bias the search of the planning tree in AlphaZero. Nonetheless, these algorithms still struggle with control problems with sparse rewards, such as goal-directed domains, where a positive reward is awarded only when reaching a goal state. In this work, we recast AlphaZero with Hindsight Experience Replay to tackle complex goal-directed planning tasks. We perform a thorough empirical evaluation in several simulated domains, including a novel application to a quantum compiling domain.

Hybrid Random Features

- Krzysztof Marcin Choromanski, Han Lin, Haoxian Chen, Arijit Sehanobish, Yuanzhe Ma, Deepali Jain, Jake Varley, Andy Zeng, Michael S Ryoo, Valerii Likhosherstov, Dmitry Kalashnikov, Vikas Sindhwani, Adrian Weller
- abstract@open-review(Poster): We propose a new class of random feature methods for linearizing softmax and Gaussian kernels called hybrid random features (HRFs) that automatically adapt the quality of kernel estimation to provide most accurate approximation in the defined regions of interest. Special instantiations of HRFs lead to well-known methods such as trigonometric (Rahimi & Recht, 2007) or (recently introduced in the context of linear-attention Transformers) positive random features (Choromanski et al., 2021). By generalizing Bochnerâ€™s Theorem for softmax/Gaussian kernels and leveraging random features for compositional kernels, the HRF-mechanism provides strong theoretical guarantees - unbiased approximation and strictly smaller worst-case relative errors than its counterparts. We conduct exhaustive empirical evaluation of HRF ranging from pointwise kernel estimation experiments, through tests on data admitting clustering structure to benchmarking implicit-attention Transformers (also for downstream Robotics applications), demonstrating its quality in a wide spectrum of machine learning problems.

Pretrained Language Model in Continual Learning: A Comparative Study

- Tongtong Wu, Massimo Caccia, Zhuang Li, Yuan-Fang Li, Guilin Qi, Gholamreza Haffari
- abstract@open-review(Poster): Continual learning (CL) is a setting in which a model learns from a stream of incoming data while avoiding to forget previously learned knowledge. Pre-trained language models (PLMs) have been successfully employed in continual learning of different natural language problems. With the rapid development of many continual learning methods and PLMs, understanding and disentangling their interactions become essential for continued improvement of continual learning performance. In this paper, we thoroughly compare the continual learning performance over the combination of 5 PLMs and 4 CL approaches on 3 benchmarks in 2 typical incremental settings. Our extensive experimental analyses reveal interesting performance differences across PLMs and across CL methods. Furthermore, our representativeness probing analyses dissect PLMsâ€™ performance characteristics in a layer-wise and task-wise manner, uncovering the extent to which their inner layers suffer from forgetting, and the effect of different CL approaches on each layer. Finally, our observations and analyses open up a number of important research questions that will inform and guide the design of effective continual learning techniques.

Salient ImageNet: How to discover spurious features in Deep Learning?

- Sahil Singla, Soheil Feizi
- abstract@open-review(Poster): Deep neural networks can be unreliable in the real world especially when they heavily use {\it spurious} features for their predictions. Focusing on image classifications, we define {\it core features} as the set of visual features that are always a part of the object definition while {\it spurious features} are the ones that are likely to {\it co-occur} with the object but not a part of it (e.g., attribute `fingers" for classband aid"). Traditional methods for discovering spurious features either require extensive human annotations (thus, not scalable), or are useful on specific models. In

this work, we introduce a {\it general} framework to discover a subset of spurious and core visual features used in inferences of a general model and localize them on a large number of images with minimal human supervision. Our methodology is based on this key idea: to identify spurious or core \textit{visual features} used in model predictions, we identify spurious or core \textit{neural features} (penultimate layer neurons of a robust model) via limited human supervision (e.g., using top 5 activating images per feature). We then show that these neural feature annotations {\it generalize} extremely well to many more images {\it without} any human supervision. We use the activation maps for these neural features as the soft masks to highlight spurious or core visual features. Using this methodology, we introduce the {\it Salient Imagenet} dataset containing core and spurious masks for a large set of samples from Imagenet. Using this dataset, we show that several popular Imagenet models rely heavily on various spurious features in their predictions, indicating the standard accuracy alone is not sufficient to fully assess model' performance specially in safety-critical applications. Code is available at \url{https://github.com/singlasahil14/salient_imagenet}.

Differentiable DAG Sampling

- Bertrand Charpentier, Simon Kibler, Stephan GÃ¼nnemann
- abstract@open-review(Poster): We propose a new differentiable probabilistic model over DAGs (DP-DAG). DP-DAG allows fast and differentiable DAG sampling suited to continuous optimization. To this end, DP-DAG samples a DAG by successively (1) sampling a linear ordering of the node and (2) sampling edges consistent with the sampled linear ordering. We further propose VI-DP-DAG, a new method for DAG learning from observational data which combines DP-DAG with variational inference. Hence,VI-DP-DAG approximates the posterior probability over DAG edges given the observed data. VI-DP-DAG is guaranteed to output a valid DAG at any time during training and does not require any complex augmented Lagrangian optimization scheme in contrast to existing differentiable DAG learning approaches. In our extensive experiments, we compare VI-DP-DAG to other differentiable DAG learning baselines on synthetic and real datasets. VI-DP-DAG significantly improves DAG structure and causal mechanism learning while training faster than competitors.

Evaluating Model-Based Planning and Planner Amortization for Continuous Control

- Arunkumar Byravan, Leonard Hasenclever, Piotr Trochim, Mehdi Mirza, Alessandro Davide Ialongo, Yuval Tassa, Jost Tobias Springenberg, Abbas Abdolmaleki, Nicolas Heess, Josh Merel, Martin Riedmiller
- abstract@open-review(Poster): There is a widespread intuition that model-based control methods should be able to surpass the data efficiency of model-free approaches. In this paper we attempt to evaluate this intuition on various challenging locomotion tasks. We take a hybrid approach, combining model predictive control (MPC) with a learned model and model-free policy learning; the learned policy serves as a proposal for MPC. We show that MPC with learned proposals and models (trained on the fly or transferred from related tasks) can significantly improve performance and data efficiency with respect to model-free methods. However, we find that well-tuned model-free agents are strong baselines even for high DoF control problems. Finally, we show that it is possible to distil a model-based planner into a policy that amortizes the planning computation without any loss of performance.

Hierarchical Few-Shot Imitation with Skill Transition Models

- Kourosh Hakhamaneshi, Ruihan Zhao, Albert Zhan, Pieter Abbeel, Michael Laskin
- abstract@open-review(Poster): A desirable property of autonomous agents is the ability to both solve long-horizon problems and generalize to unseen tasks. Recent advances in data-driven skill learning have shown that extracting behavioral priors from offline data can enable agents to solve challenging long-horizon tasks with reinforcement learning. However, generalization to tasks unseen during behavioral prior training remains an outstanding challenge. To this end, we present Few-shot Imitation with Skill Transition Models (FIST), an algorithm that extracts skills from offline data and utilizes them to generalize to unseen tasks given a few downstream demonstrations. FIST learns an inverse skill dynamics model, a distance function, and utilizes a semi-parametric approach for imitation. We show that FIST is capable of generalizing to new tasks and substantially outperforms prior baselines in navigation experiments requiring traversing unseen parts of a large maze and 7-DoF robotic arm experiments requiring manipulating previously unseen objects in a kitchen.

End-to-End Learning of Probabilistic Hierarchies on Graphs

- Daniel ZÃ¼gner, Bertrand Charpentier, Morgane Ayle, Sascha Geringer, Stephan GÃ¼nnemann
- abstract@open-review(Poster): We propose a novel probabilistic model over hierarchies on graphs obtained by continuous relaxation of tree-based hierarchies. We draw connections to Markov chain theory, enabling us to perform hierarchical clustering by efficient end-to-end optimization of relaxed versions of quality metrics such as Dasgupta cost or Tree-Sampling Divergence (TSD). We show that our model learns rich, high-quality hierarchies present in 11 real world graphs, including a large graph with 2.3M nodes. Our model consistently outperforms recent as well as strong traditional baselines such as average linkage. Our model also obtains strong results on link prediction despite not being trained on this task, highlighting the quality of the hierarchies discovered by our model.

GeneDisco: A Benchmark for Experimental Design in Drug Discovery

- Arash Mehrjou, Ashkan Soleymani, Andrew Jesson, Pascal Notin, Yarin Gal, Stefan Bauer, Patrick Schwab
- abstract@open-review(Poster): In vitro cellular experimentation with genetic interventions, using for example CRISPR technologies, is an essential step in early-stage drug discovery and target validation that serves to assess initial hypotheses about causal associations between biological mechanisms and disease pathologies. With billions of potential hypotheses to test, the experimental design space for in vitro genetic experiments is extremely vast, and the available experimental capacity - even at the largest research institutions in the world - pales in relation to the size of this biological hypothesis space. Machine learning methods, such as active and reinforcement learning, could aid in optimally exploring the vast biological space by integrating prior knowledge from various information sources as well as extrapolating to yet unexplored areas of the experimental design space based on available data. However, there exist no standardised benchmarks and data sets for this challenging task and little research has been conducted in this area to date. Here, we introduce GeneDisco, a benchmark suite for evaluating active learning algorithms for experimental design in drug discovery. GeneDisco contains a curated set of multiple publicly available experimental data sets as well as open-source implementations of state-of-the-art active learning policies for experimental design and exploration.

GraphENS: Neighbor-Aware Ego Network Synthesis for Class-Imbalanced Node Classification

- Joonhyung Park, Jaeyun Song, Eunho Yang
- abstract@open-review(Poster): In many real-world node classification scenarios, nodes are highly class-imbalanced, where graph neural networks (GNNs) can be readily biased to major class instances. Albeit existing class imbalance approaches in other domains can alleviate this issue to some extent, they do not consider the impact of message passing between nodes. In this paper, we hypothesize that overfitting to the neighbor sets of minor class due to message passing is a major challenge for class-imbalanced node classification. To tackle this issue, we propose GraphENS, a novel augmentation method that synthesizes the whole ego network for minor class (minor node and its one-hop neighbors) by combining two different ego networks based on their similarity. Additionally, we introduce a saliency-based node mixing method to exploit the abundant class-generic attributes of other nodes while blocking the injection of class-specific features. Our approach consistently outperforms the baselines over multiple node classification benchmark datasets and architectures.

Continuously Discovering Novel Strategies via Reward-Switching Policy Optimization

- Zihan Zhou, Wei Fu, Bingliang Zhang, Yi Wu
- abstract@open-review(Poster): We present Reward-Switching Policy Optimization (RSPO), a paradigm to discover diverse strategies in complex RL environments by iteratively finding novel policies that are both locally optimal and sufficiently different from existing ones. To encourage the learning policy to consistently converge towards a previously undiscovered local optimum, RSPO switches between extrinsic and intrinsic rewards via a trajectory-based novelty measurement during the optimization process. When a sampled trajectory is sufficiently distinct, RSPO performs standard policy optimization with extrinsic rewards. For trajectories with high likelihood under existing policies, RSPO utilizes an intrinsic diversity reward to promote exploration. Experiments show that RSPO is able to discover a wide spectrum of strategies in a variety of domains, ranging from single-agent navigation tasks and MuJoCo control to multi-agent stag-hunt games and the StarCraft II Multi-Agent Challenge.

Learning to Remember Patterns: Pattern Matching Memory Networks for Traffic Forecasting

- Hyunwook Lee, Seungmin Jin, Hyeshin Chu, Hongkyu Lim, Sungahn Ko
- abstract@open-review(Poster): Traffic forecasting is a challenging problem due to complex road networks and sudden speed changes caused by various events on roads. Several models have been proposed to solve this challenging problem, with a focus on learning the spatio-temporal dependencies of roads. In this work, we propose a new perspective for converting the forecasting problem into a pattern-matching task, assuming that large traffic data can be represented by a set of patterns. To evaluate the validity of this new perspective, we design a novel traffic forecasting model called Pattern-Matching Memory Networks (PM-MemNet), which learns to match input data to representative patterns with a key-value memory structure. We first extract and cluster representative traffic patterns that serve as keys in the memory. Then, by matching the extracted keys and inputs, PM-MemNet acquires the necessary information on existing traffic patterns from the memory and uses it for forecasting. To model the spatio-temporal correlation of traffic, we proposed a novel memory architecture, GCMem, which integrates attention and graph convolution. The experimental results indicate that PM-MemNet is more accurate than state-of-the-art models, such as Graph WaveNet, with higher responsiveness. We also present a qualitative analysis describing how PM-MemNet works and achieves higher accuracy when road speed changes rapidly.

Why Propagate Alone? Parallel Use of Labels and Features on Graphs

- Yangkun Wang, Jiarui Jin, Weinan Zhang, Yang Yongyi, Jiuhai Chen, Quan Gan, Yong Yu, Zheng Zhang, Zengfeng Huang, David Wipf
- abstract@open-review(Poster): One of the challenges of graph-based semi-supervised learning over ordinary supervised learning for classification tasks lies in label utilization. The direct use of ground-truth labels in graphs for training purposes can result in a parametric model learning trivial degenerate solutions (e.g., an identity mapping from input to output). In addressing this issue, a label trick has recently been proposed in the literature and applied to a wide range of graph neural network (GNN) architectures, achieving state-of-the-art results on various datasets. The essential idea is to randomly split the observed labels on the graph and use a fraction of them as input to the model (along with original node features), and predict the remaining fraction. Despite its success in enabling GNNs to propagate features and labels simultaneously, this approach has never been analyzed from a theoretical perspective, nor fully explored across certain natural use cases. In this paper, we demonstrate that under suitable settings, this stochastic trick can be reduced to a more interpretable deterministic form, allowing us to better explain its behavior, including an emergent regularization effect, and motivate broader application scenarios. Our experimental results corroborate these analyses while also demonstrating improved node classification performance applying the label trick in new domains.

Learning by Directional Gradient Descent

- David Silver, Anirudh Goyal, Ivo Danihelka, Matteo Hessel, Hado van Hasselt
- abstract@open-review(Poster): How should state be constructed from a sequence of observations, so as to best achieve some objective? Most deep learning methods update the parameters of the state representation by gradient descent. However, no prior method for computing the gradient is fully satisfactory, for example consuming too much memory, introducing too much variance, or adding too much bias. In this work, we propose a new learning algorithm that addresses these limitations. The basic idea is to update the parameters of the representation by using the directional derivative along a candidate direction, a quantity that may be computed online with the same computational cost as the representation itself. We consider several different choices of candidate direction, including random selection and approximations to the true gradient, and investigate their performance on several synthetic tasks.

Maximum Entropy RL (Provably) Solves Some Robust RL Problems

- Benjamin Eysenbach, Sergey Levine
- abstract@open-review(Poster): Many potential applications of reinforcement learning (RL) require guarantees that the agent will perform well in the face of disturbances to the dynamics or reward function. In this paper, we prove theoretically that maximum entropy (MaxEnt) RL maximizes a lower bound on a robust RL objective, and thus can be used to learn policies that are robust to some disturbances in the dynamics and the reward function. While this capability of MaxEnt RL has been observed empirically in prior work, to the best of our knowledge our work provides the first rigorous proof and theoretical characterization of the MaxEnt RL robust set. While a number of prior robust RL algorithms have been designed to handle similar disturbances to the reward function or dynamics, these methods typically require additional moving parts and hyperparameters on top of a base RL algorithm. In contrast, our results suggest that MaxEnt RL by itself is robust to certain disturbances, without requiring any additional modifications. While this does not imply that MaxEnt RL is the best available robust RL method, MaxEnt RL is a simple robust RL method with appealing formal guarantees.

A Unified Contrastive Energy-based Model for Understanding the Generative Ability of Adversarial Training

- Yifei Wang, Yisen Wang, Jiansheng Yang, Zhouchen Lin
- abstract@open-review(Poster): Adversarial Training (AT) is known as an effective approach to enhance the robustness of deep neural networks. Recently researchers notice that robust models with AT have good generative ability and can synthesize realistic images, while the reason behind it is yet under-explored. In this paper, we demystify this phenomenon by developing a unified probabilistic framework, called Contrastive Energy-based Models (CEM). On the one hand, we provide the first probabilistic characterization of AT through a unified understanding of robustness and generative ability. On the other hand, our unified framework can be extended to the unsupervised scenario, which interprets unsupervised contrastive learning as an important sampling of CEM. Based on these, we propose a principled method to develop adversarial learning and sampling methods. Experiments show that the sampling methods derived from our framework improve the sample quality in both supervised and unsupervised learning. Notably, our unsupervised adversarial sampling method achieves an Inception score of 9.61 on CIFAR-10, which is superior to previous energy-based models and comparable to state-of-the-art generative models.

Equivariant and Stable Positional Encoding for More Powerful Graph Neural Networks

- Haorui Wang, Haoteng Yin, Muhan Zhang, Pan Li
- abstract@open-review(Poster): Graph neural networks (GNN) have shown great advantages in many graph-based learning tasks but often fail to predict accurately for a task-based on sets of nodes such as link/motif prediction and so on. Many works have recently proposed to address this problem by using random node features or node distance features. However, they suffer from either slow convergence, inaccurate prediction, or high complexity. In this

work, we revisit GNNs that allow using positional features of nodes given by positional encoding (PE) techniques such as Laplacian Eigenmap, Deepwalk, etc. GNNs with PE often get criticized because they are not generalizable to unseen graphs (inductive) or stable. Here, we study these issues in a principled way and propose a provable solution, a class of GNN layers termed PEG with rigorous mathematical analysis. PEG uses separate channels to update the original node features and positional features. PEG imposes permutation equivariance w.r.t. the original node features and rotation equivariance w.r.t. the positional features simultaneously. Extensive link prediction experiments over 8 real-world networks demonstrate the advantages of PEG in generalization and scalability. Code is available at https://github.com/Graph-COM/PEG.

BadPre: Task-agnostic Backdoor Attacks to Pre-trained NLP Foundation Models

- Kangjie Chen, Yuxian Meng, Xiaofei Sun, Shangwei Guo, Tianwei Zhang, Jiwei Li, Chun Fan
- abstract@open-review(Poster): Pre-trained Natural Language Processing (NLP) models, which can be adapted to a variety of downstream language tasks via fine-tuning, highly accelerate the learning progress of NLP models. However, NLP models have been shown to be vulnerable to backdoor attacks. Previous NLP backdoor attacks mainly focus on one specific task. This limitation makes existing solutions less applicable to different NLP models which have been widely used in various tasks. In this work, we propose BadPre, the first backdoor attack against various downstream models built based on pre-trained NLP models. BadPre can launch trojan attacks against different language tasks with the same trigger. The key insight of our approach is that downstream models can inherit the security characteristics from the pre-trained models. Specifically, we leverage data posing to the pre-trained NLP models and then inference the downstream models with sentences embedded triggers. Furthermore, to fool backdoor detectors, we design a novel adversarial attack method to generate a more robust trigger. Experimental results indicate that our approach can effectively attack a wide range of downstream NLP tasks and exhibit significant robustness against backdoor detectors.

Shallow and Deep Networks are Near-Optimal Approximators of Korobov Functions

- Moise Blanchard, Mohammed Amine Bennouna
- abstract@open-review(Poster): In this paper, we analyze the number of neurons and training parameters that a neural network needs to approximate multivariate functions of bounded second mixed derivatives --- Korobov functions. We prove upper bounds on these quantities for shallow and deep neural networks, drastically lessening the curse of dimensionality. Our bounds hold for general activation functions, including ReLU. We further prove that these bounds nearly match the minimal number of parameters any continuous function approximator needs to approximate Korobov functions, showing that neural networks are near-optimal function approximators.

What Makes Better Augmentation Strategies? Augment Difficult but Not too Different

- Jaehyung Kim, Dongyeop Kang, Sungsoo Ahn, Jinwoo Shin
- abstract@open-review(Poster): The practice of data augmentation has been extensively used to boost the performance of deep neural networks for various NLP tasks. It is more effective when only a limited number of labeled samples is available, e.g., low-data or class-imbalanced regimes. Most current augmentation techniques rely on parameter tuning or inherent randomness; hence, their effectiveness largely varies on the tasks. To efficiently find the best augmentation strategy for each task, learning data augmentation policy is a promising solution, but the question of what makes a good augmentation in NLP tasks and how to design the reward function for learning a good policy remains under-explored. To answer this, we hypothesize that good data augmentation should construct more diverse and challenging samples for providing informative training signals, while avoiding the risk of losing the semantics of original samples. Therefore, we design a novel reward function for updating the augmentation policy to construct difficult but not too different samples (DND). Particularly, we jointly optimize a data augmentation policy while training the model, to construct the augmented samples with low confidence but a high semantic similarity with original ones. In addition, we introduce a sample re-weighting scheme to focus on difficult augmented samples after the original ones are learned confidently for more effective learning from the augmented ones. Our learning-based augmentation outperforms the recent state-of-the-art augmentation schemes on various text classification tasks and GLUE benchmark by successfully discovering the effective augmentations for each task. Remarkably, our method is more effective on the challenging low-data and class-imbalanced regimes, and the learned augmentation policy is well-transferable to the different tasks and models.

Generative Pseudo-Inverse Memory

- Kha Pham, Hung Le, Man Ngo, Truyen Tran, Bao Ho, Svetha Venkatesh
- abstract@open-review(Poster): We propose Generative Pseudo-Inverse Memory (GPM), a class of deep generative memory models that are fast to write in and read out. Memory operations are recast as seeking robust solutions of linear systems, which naturally lead to the use of matrix pseudo-inverses. The pseudo-inverses are iteratively approximated, with practical computation complexity of almost $O(1)$. We prove theoretically and verify empirically that our model can retrieve exactly what have been written to the memory under mild conditions. A key capability of GPM is iterative reading, during which the attractor dynamics towards fixed points are enabled, allowing the model to iteratively improve sample quality in denoising and generating. More impressively, GPM can store a large amount of data while maintaining key abilities of accurate retrieving of stored patterns, denoising of corrupted data and generating novel samples. Empirically we demonstrate the efficiency and versatility of GPM on a comprehensive suite of experiments involving binarized MNIST, binarized Omniglot, FashionMNIST, CIFAR10 & CIFAR100 and CelebA.

A Deep Variational Approach to Clustering Survival Data

- Laura Manduchi, RiÄ ards MarcinkeviÄ s, Michela C. Massi, Thomas Weikert, Alexander Sauter, Verena Gotta, Timothy MÃ¼ller, Flavio Vasella, Marian C. Neidert, Marc Pfister, Bram Stieltjes, Julia E Vogt
- abstract@open-review(Poster): In this work, we study the problem of clustering survival data â€” a challenging and so far under-explored task. We introduce a novel semi-supervised probabilistic approach to cluster survival data by leveraging recent advances in stochastic gradient variational inference. In contrast to previous work, our proposed method employs a deep generative model to uncover the underlying distribution of both the explanatory variables and censored survival times. We compare our model to the related work on clustering and mixture models for survival data in comprehensive experiments on a wide range of synthetic, semi-synthetic, and real-world datasets, including medical imaging data. Our method performs better at identifying clusters and is competitive at predicting survival times. Relying on novel generative assumptions, the proposed model offers a holistic perspective on clustering survival data and holds a promise of discovering subpopulations whose survival is regulated by different generative mechanisms.

GPT-Critic: Offline Reinforcement Learning for End-to-End Task-Oriented Dialogue Systems

- Youngsoo Jang, Jongmin Lee, Kee-Eung Kim
- abstract@open-review(Poster): Training a task-oriented dialogue agent can be naturally formulated as offline reinforcement learning (RL) problem, where the agent aims to learn a conversational strategy to achieve user goals, only from a dialogue corpus. It is very challenging in terms of RL since the natural language action space is astronomical, while feasible (syntactically and semantically correct) actions are very sparse. Thus, standard RL methods easily fail and generate responses diverging from human language, even when fine-tuning a powerful pre-trained language model. In this paper, we introduce GPT-Critic, an offline RL method for task-oriented dialogue. GPT-Critic is built upon GPT-2, fine-tuning the language model through behavior cloning of the critic-guided self-generated sentences. GPT-Critic is essentially free from the issue of diverging from human language since it learns from the sentences sampled from the pre-trained language model. In the experiments, we demonstrate that our algorithm outperforms the state-of-the-art in the task-oriented dialogue benchmarks including MultiWOZ 2.0 and ConvLab.

Charformer: Fast Character Transformers via Gradient-based Subword Tokenization

- Yi Tay, Vinh Q. Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, Donald Metzler
- abstract@open-review(Poster): State-of-the-art models in natural language processing rely on separate rigid subword tokenization algorithms, which limit their generalization ability and adaptation to new settings. In this paper, we propose a new model inductive bias that learns a subword tokenization end-to-end as part of the model. To this end, we introduce a soft gradient-based subword tokenization module (GBST) that automatically learns latent subword representations from characters in a data-driven fashion. Concretely, GBST enumerates candidate subword blocks and learns to score them in a position-wise fashion using a block scoring network. We additionally introduce Charformer, a deep Transformer model that integrates GBST and operates on the character level. Via extensive experiments on English GLUE, multilingual, and noisy text datasets, we show that Charformer outperforms a series of competitive character-level baselines while generally performing on par and sometimes outperforming subword-based models. Additionally, Charformer is fast, improving the speed of vanilla character-level Transformers by up to while maintaining quality. We believe this work paves the way for highly performant token-free models that are trained completely end-to-end.

Regularized Autoencoders for Isometric Representation Learning

- Yonghyeon LEE, Sangwoong Yoon, MinJun Son, Frank C. Park
- abstract@open-review(Poster): The recent success of autoencoders for representation learning can be traced in large part to the addition of a regularization term. Such regularized autoencoders ``constrain" the representation so as to prevent overfitting to the data while producing a parsimonious generative model. A regularized autoencoder should in principle learn not only the data manifold, but also a set of geometry-preserving coordinates for the latent representation space; by geometry-preserving we mean that the latent space representation should attempt to preserve actual distances and angles on the data manifold. In this paper we first formulate a hierarchy for geometry-preserving mappings (isometry, conformal mapping of degree k, area-preserving mappings). We then show that a conformal regularization term of degree zero -- i.e., one that attempts to preserve angles and relative distances, instead of angles and exact distances -- produces data representations that are superior to other existing methods. Applying our algorithm to an unsupervised information retrieval task for CelebA data with 40 annotations, we achieve 79\% precision at five retrieved images, an improvement of more than 10\% compared to recent related work. Code is available at https://github.com/Gabe-YHLee/IRVAE-public.

Knowledge Removal in Sampling-based Bayesian Inference

- Shaopeng Fu, Fengxiang He, Dacheng Tao
- abstract@open-review(Poster): The right to be forgotten has been legislated in many countries, but its enforcement in the AI industry would cause unbearable costs. When single data deletion requests come, companies may need to delete the whole models learned with massive resources. Existing works propose methods to remove knowledge learned from data for explicitly parameterized models, which however are not appliable to the sampling-based Bayesian inference, {\it i.e.}, Markov chain Monte Carlo (MCMC), as MCMC can only infer implicit distributions. In this paper, we propose the first machine unlearning algorithm for MCMC. We first convert the MCMC unlearning problem into an explicit optimization problem. Based on this problem conversion, an {\it MCMC influence function} is designed to provably characterize the learned knowledge from data, which then delivers the MCMC unlearning algorithm. Theoretical analysis shows that MCMC unlearning would not compromise the generalizability of the MCMC models. Experiments on Gaussian mixture models and Bayesian neural networks confirm the effectiveness of the proposed algorithm. The code is available at \url{https://github.com/fshp971/mcmc-unlearning}.

Actor-critic is implicitly biased towards high entropy optimal policies

- Yuzheng Hu, Ziwei Ji, Matus Telgarsky
- abstract@open-review(Poster): We show that the simplest actor-critic method â€” a linear softmax policy updated with TD through interaction with a linear MDP, but featuring no explicit regularization or exploration â€” does not merely find an optimal policy, but moreover prefers high entropy optimal policies. To demonstrate the strength of this bias, the algorithm not only has no regularization, no projections, and no exploration like ϵ-greedy, but is moreover trained on a single trajectory with no resets. The key consequence of the high entropy bias is that uniform mixing assumptions on the MDP, which exist in some form in all prior work, can be dropped: the implicit regularization of the high entropy bias is enough to ensure that all chains mix and an optimal policy is reached with high probability. As auxiliary contributions, this work decouples concerns between the actor and critic by writing the actor update as an explicit mirror descent, provides tools to uniformly bound mixing times within KL balls of policy space, and provides a projection-free TD analysis with its own implicit bias which can be run from an unmixed starting distribution.

Igeood: An Information Geometry Approach to Out-of-Distribution Detection

- Eduardo Dadalto Camara Gomes, Florence Alberge, Pierre Duhamel, Pablo Piantanida
- abstract@open-review(Poster): Reliable out-of-distribution (OOD) detection is fundamental to implementing safer modern machine learning (ML) systems. In this paper, we introduce Igeood, an effective method for detecting OOD samples. Igeood applies to any pre-trained neural network, works under various degrees of access to the ML model, does not require OOD samples or assumptions on the OOD data but can also benefit (if available) from OOD samples. By building on the geodesic (Fisher-Rao) distance between the underlying data distributions, our discriminator can combine confidence scores from the logits outputs and the learned features of a deep neural network. Empirically, we show that Igeood outperforms competing state-of-the-art methods on a variety of network architectures and datasets.

Bag of Instances Aggregation Boosts Self-supervised Distillation

- Haohang Xu, Jiemin Fang, XIAOPENG ZHANG, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, Qi Tian
- abstract@open-review(Poster): Recent advances in self-supervised learning have experienced remarkable progress, especially for contrastive learning based methods, which regard each image as well as its augmentations as an individual class and try to distinguish them from all other images. However, due to the large quantity of exemplars, this kind of pretext task intrinsically suffers from slow convergence and is hard for optimization. This is especially true for small-scale models, in which we find the performance drops dramatically comparing with its supervised counterpart. In this paper, we propose a simple but effective distillation strategy for unsupervised learning. The highlight is that the relationship among similar samples counts and can be seamlessly transferred to the student to boost the performance. Our method, termed as BINGO, which is short for Bag of InstaNces aGgregatiOn, targets at transferring the relationship learned by the teacher to the student. Here bag of instances indicates a set of similar samples constructed by the teacher and are grouped within a bag, and the goal of distillation is to aggregate compact representations over the student with respect to instances in a bag. Notably, BINGO achieves new state-of-the-art performance on small-scale models, i.e., 65.5% and 68.9% top-1 accuracies with linear evaluation on ImageNet, using ResNet-18 and ResNet-34 as the backbones respectively, surpassing baselines (52.5% and 57.4% top-1 accuracies) by a significant margin. The code is available at https://github.com/haohang96/bingo.

Stability Regularization for Discrete Representation Learning

- Adeel Pervez, Efstratios Gavves
- abstract@open-review(Poster): We present a method for training neural network models with discrete stochastic variables. The core of the method is \emph{stability regularization}, which is a regularization procedure based on the idea of noise stability developed in Gaussian isoperimetric theory in the

analysis of Gaussian functions. Stability regularization is method to make the output of continuous functions of Gaussian random variables close to discrete, that is binary or categorical, without the need for significant manual tuning. The method allows control over the extent to which a Gaussian function's output is close to discrete, thus allowing for continued flow of gradient. The method can be used standalone or in combination with existing continuous relaxation methods. We validate the method in a broad range of experiments using discrete variables including neural relational inference, generative modeling, clustering and conditional computing.

Unrolling PALM for Sparse Semi-Blind Source Separation

- Mohammad Fahes, Christophe Kervazo, JÃ©rÃ´me Bobin, Florence Tupin
- abstract@open-review(Poster): Sparse Blind Source Separation (BSS) has become a well established tool for a wide range of applications â€“ for instance, in astrophysics and remote sensing. Classical sparse BSS methods, such as the Proximal Alternating Linearized Minimization (PALM) algorithm, nevertheless often suffer from a difficult hyper-parameter choice, which undermines their results. To bypass this pitfall, we propose in this work to build on the thriving field of algorithm unfolding/unrolling. Unrolling PALM enables to leverage the data-driven knowledge stemming from realistic simulations or ground-truth data by learning both PALM hyper-parameters and variables. In contrast to most existing unrolled algorithms, which assume a fixed known dictionary during the training and testing phases, this article further emphasizes on the ability to deal with variable mixing matrices (a.k.a. dictionaries). The proposed Learned PALM (LPALM) algorithm thus enables to perform semi-blind source separation, which is key to increase the generalization of the learnt model in real-world applications. We illustrate the relevance of LPALM in astrophysical multispectral imaging: the algorithm not only needs up to $10^4â^’10^5$ times less iterations than PALM, but also improves the separation quality, while avoiding the cumbersome hyper-parameter and initialization choice of PALM. We further show that LPALM outperforms other unrolled source separation methods in the semi-blind setting.

Fast Generic Interaction Detection for Model Interpretability and Compression

- Tianjian Zhang, Feng Yin, Zhi-Quan Luo
- abstract@open-review(Poster): The ability of discovering feature interactions in a black-box model is vital to explainable deep learning. We propose a principled, global interaction detection method by casting our target as a multi-arm bandits problem and solving it swiftly with the UCB algorithm. This adaptive method is free of ad-hoc assumptions and among the cutting-edge methods with outstanding detection accuracy and stability. Based on the detection outcome, a lightweight and interpretable deep learning model (called ParaACE) is further built using the alternating conditional expectation (ACE) method. Our proposed ParaACE improves the prediction performance by 26 % and reduces the model size by 100+ times as compared to its Teacher model over various datasets. Furthermore, we show the great potential of our method for scientific discovery through interpreting various real datasets in the economics and smart medicine sectors. The code is available at https://github.com/zhangtj1996/ParaACE.

Reversible Instance Normalization for Accurate Time-Series Forecasting against Distribution Shift

- Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, Jaegul Choo
- abstract@open-review(Poster): Statistical properties such as mean and variance often change over time in time series, i.e., time-series data suffer from a distribution shift problem. This change in temporal distribution is one of the main challenges that prevent accurate time-series forecasting. To address this issue, we propose a simple yet effective normalization method called reversible instance normalization (RevIN), a generally-applicable normalization-and-denormalization method with learnable affine transformation. The proposed method is symmetrically structured to remove and restore the statistical information of a time-series instance, leading to significant performance improvements in time-series forecasting, as shown in Fig. 1. We demonstrate the effectiveness of RevIN via extensive quantitative and qualitative analyses on various real-world datasets, addressing the distribution shift problem.

On the Pitfalls of Analyzing Individual Neurons in Language Models

- Omer Antverg, Yonatan Belinkov
- abstract@open-review(Poster): While many studies have shown that linguistic information is encoded in hidden word representations, few have studied individual neurons, to show how and in which neurons it is encoded. Among these, the common approach is to use an external probe to rank neurons according to their relevance to some linguistic attribute, and to evaluate the obtained ranking using the same probe that produced it. We show two pitfalls in this methodology:
    1. It confounds distinct factors: probe quality and ranking quality. We separate them and draw conclusions on each.
    2. It focuses on encoded information, rather than information that is used by the model. We show that these are not the same. We compare two recent ranking methods and a simple one we introduce, and evaluate them with regard to both of these aspects.

Query Embedding on Hyper-Relational Knowledge Graphs

- Dimitrios Alivanistos, Max Berrendorf, Michael Cochez, Mikhail Galkin
- abstract@open-review(Poster): Multi-hop logical reasoning is an established problem in the field of representation learning on knowledge graphs (KGs). It subsumes both one-hop link prediction as well as other more complex types of logical queries. Existing algorithms operate only on classical, triple-based graphs, whereas modern KGs often employ a hyper-relational modeling paradigm. In this paradigm, typed edges may have several key-value pairs known as qualifiers that provide fine-grained context for facts. In queries, this context modifies the meaning of relations, and usually reduces the answer set. Hyper-relational queries are often observed in real-world KG applications, and existing approaches for approximate query answering cannot make use of qualifier pairs. In this work, we bridge this gap and extend the multi-hop reasoning problem to hyper-relational KGs allowing to tackle this new type of complex queries. Building upon recent advancements in Graph Neural Networks and query embedding techniques, we study how to embed and answer hyper-relational conjunctive queries. Besides that, we propose a method to answer such queries and demonstrate in our experiments that qualifiers improve query answering on a diverse set of query patterns.

Neural Solvers for Fast and Accurate Numerical Optimal Control

- Federico Berto, Stefano Massaroli, Michael Poli, Jinkyoo Park
- abstract@open-review(Poster): Synthesizing optimal controllers for dynamical systems often involves solving optimization problems with hard real-time constraints. These constraints determine the class of numerical methods that can be applied: computationally expensive but accurate numerical routines are replaced by fast and inaccurate methods, trading inference time for solution accuracy. This paper provides techniques to improve the quality of optimized control policies given a fixed computational budget. We achieve the above via a hypersolvers approach, which hybridizes a differential equation solver and a neural network. The performance is evaluated in direct and receding-horizon optimal control tasks in both low and high dimensions, where the proposed approach shows consistent Pareto improvements in solution accuracy and control performance.

PSA-GAN: Progressive Self Attention GANs for Synthetic Time Series

- Paul Jeha, Michael Bohlke-Schneider, Pedro Mercado, Shubham Kapoor, Rajbir Singh Nirwan, Valentin Flunkert, Jan Gasthaus, Tim Januschowski
- abstract@open-review(Poster): Realistic synthetic time series data of sufficient length enables practical applications in time series modeling tasks, such as forecasting, but remains a challenge. In this paper we present PSA-GAN, a generative adversarial network (GAN) that generates long time series samples

of high quality using progressive growing of GANs and self-attention. We show that PSA-GAN can be used to reduce the error in several downstream forecasting tasks over baselines that only use real data. We also introduce a Frechet-Inception Distance-like score for time series, Context-FID, assessing the quality of synthetic time series samples. We find that Context-FID is indicative for downstream performance. Therefore, Context-FID could be a useful tool to develop time series GAN models.

ToM2C: Target-oriented Multi-agent Communication and Cooperation with Theory of Mind

- Yuanfei Wang, fangwei zhong, Jing Xu, Yizhou Wang
- abstract@open-review(Poster): Being able to predict the mental states of others is a key factor to effective social interaction. It is also crucial for distributed multi-agent systems, where agents are required to communicate and cooperate. In this paper, we introduce such an important social-cognitive skill, i.e. Theory of Mind (ToM), to build socially intelligent agents who are able to communicate and cooperate effectively to accomplish challenging tasks. With ToM, each agent is capable of inferring the mental states and intentions of others according to its (local) observation. Based on the inferred states, the agents decide "when" and with "whom" to share their intentions. With the information observed, inferred, and received, the agents decide their sub-goals and reach a consensus among the team. In the end, the low-level executors independently take primitive actions to accomplish the sub-goals. We demonstrate the idea in two typical target-oriented multi-agent tasks: cooperative navigation and multi-sensor target coverage. The experiments show that the proposed model not only outperforms the state-of-the-art methods on reward and communication efficiency, but also shows good generalization across different scales of the environment.

Better Supervisory Signals by Observing Learning Paths

- Yi Ren, Shangmin Guo, Danica J. Sutherland
- abstract@open-review(Poster): Better-supervised models might have better performance. In this paper, we first clarify what makes for good supervision for a classification problem, and then explain two existing label refining methods, label smoothing and knowledge distillation, in terms of our proposed criterion. To further answer why and how better supervision emerges, we observe the learning path, i.e., the trajectory of the model's predictions during training, for each training sample. We find that the model can spontaneously refine "bad" labels through a "zig-zag" learning path, which occurs on both toy and real datasets. Observing the learning path not only provides a new perspective for understanding knowledge distillation, overfitting, and learning dynamics, but also reveals that the supervisory signal of a teacher network can be very unstable near the best points in training on real tasks. Inspired by this, we propose a new knowledge distillation scheme, Filter-KD, which improves downstream classification performance in various settings.

TAda! Temporally-Adaptive Convolutions for Video Understanding

- Ziyuan Huang, Shiwei Zhang, Liang Pan, Zhiwu Qing, Mingqian Tang, Ziwei Liu, Marcelo H Ang Jr
- abstract@open-review(Poster): Spatial convolutions are widely used in numerous deep video models. It fundamentally assumes spatio-temporal invariance, i.e., using shared weights for every location in different frames. This work presents Temporally-Adaptive Convolutions (TAdaConv) for video understanding, which shows that adaptive weight calibration along the temporal dimension is an efficient way to facilitate modelling complex temporal dynamics in videos. Specifically, TAdaConv empowers the spatial convolutions with temporal modelling abilities by calibrating the convolution weights for each frame according to its local and global temporal context. Compared to previous temporal modelling operations, TAdaConv is more efficient as it operates over the convolution kernels instead of the features, whose dimension is an order of magnitude smaller than the spatial resolutions. Further, the kernel calibration brings an increased model capacity. We construct TAda2D and TAdaConvNeXt networks by replacing the 2D convolutions in ResNet and ConvNeXt with TAdaConv, which leads to at least on par or better performance compared to state-of-the-art approaches on multiple video action recognition and localization benchmarks. We also demonstrate that as a readily plug-in operation with negligible computation overhead, TAdaConv can effectively improve many existing video models with a convincing margin.

Learning a subspace of policies for online adaptation in Reinforcement Learning

- Jean-Baptiste Gaya, Laure Soulier, Ludovic Denoyer
- abstract@open-review(Poster): Deep Reinforcement Learning (RL) is mainly studied in a setting where the training and the testing environments are similar. But in many practical applications, these environments may differ. For instance, in control systems, the robot(s) on which a policy is learned might differ from the robot(s) on which a policy will run. It can be caused by different internal factors (e.g., calibration issues, system attrition, defective modules) or also by external changes (e.g., weather conditions). There is a need to develop RL methods that generalize well to variations of the training conditions. In this article, we consider the simplest yet hard to tackle generalization setting where the test environment is unknown at train time, forcing the agent to adapt to the system's new dynamics. This online adaptation process can be computationally expensive (e.g., fine-tuning) and cannot rely on meta-RL techniques since there is just a single train environment. To do so, we propose an approach where we learn a subspace of policies within the parameter space. This subspace contains an infinite number of policies that are trained to solve the training environment while having different parameter values. As a consequence, two policies in that subspace process information differently and exhibit different behaviors when facing variations of the train environment. Our experiments carried out over a large variety of benchmarks compare our approach with baselines, including diversity-based methods. In comparison, our approach is simple to tune, does not need any extra component (e.g., discriminator) and learns policies able to gather a high reward on unseen environments.

ZeroFL: Efficient On-Device Training for Federated Learning with Local Sparsity

- Xinchi Qiu, Javier Fernandez-Marques, Pedro PB Gusmao, Yan Gao, Titouan Parcollet, Nicholas Donald Lane
- abstract@open-review(Poster): When the available hardware cannot meet the memory and compute requirements to efficiently train high performing machine learning models, a compromise in either the training quality or the model complexity is needed. In Federated Learning (FL), nodes are orders of magnitude more constrained than traditional server-grade hardware and are often battery powered, severely limiting the sophistication of models that can be trained under this paradigm. While most research has focused on designing better aggregation strategies to improve convergence rates and in alleviating the communication costs of FL, fewer efforts have been devoted to accelerating on-device training. Such stage, which repeats hundreds of times (i.e. every round) and can involve thousands of devices, accounts for the majority of the time required to train federated models and, the totality of the energy consumption at the client side. In this work, we present the first study on the unique aspects that arise when introducing sparsity at training time in FL workloads. We then propose ZeroFL, a framework that relies on highly sparse operations to accelerate on-device training. Models trained with ZeroFL and 95% sparsity achieve up to 2.3% higher accuracy compared to competitive baselines obtained from adapting a state-of-the-art sparse training framework to the FL setting.

Gaussian Mixture Convolution Networks

- Adam Celarek, Pedro Hermosilla, Bernhard Kerbl, Timo Ropinski, Michael Wimmer
- abstract@open-review(Poster): This paper proposes a novel method for deep learning based on the analytical convolution of multidimensional Gaussian mixtures. In contrast to tensors, these do not suffer from the curse of dimensionality and allow for a compact representation, as data is only stored where details exist. Convolution kernels and data are Gaussian mixtures with unconstrained weights, positions, and covariance matrices. Similar to discrete convolutional networks, each convolution step produces several feature channels, represented by independent Gaussian mixtures. Since traditional transfer functions like ReLUs do not produce Gaussian mixtures, we propose using a fitting of these functions instead. This fitting step also acts as a pooling layer

if the number of Gaussian components is reduced appropriately. We demonstrate that networks based on this architecture reach competitive accuracy on Gaussian mixtures fitted to the MNIST and ModelNet data sets.

How Does SimSiam Avoid Collapse Without Negative Samples? A Unified Understanding with Self-supervised Contrastive Learning

- Chaoning Zhang, Kang Zhang, Chenshuang Zhang, Trung X. Pham, Chang D. Yoo, In So Kweon
- abstract@open-review(Poster): To avoid collapse in self-supervised learning (SSL), a contrastive loss is widely used but often requires a large number of negative samples. Without negative samples yet achieving competitive performance, a recent work~\citep{chen2021exploring} has attracted significant attention for providing a minimalist simple Siamese (SimSiam) method to avoid collapse. However, the reason for how it avoids collapse without negative samples remains not fully clear and our investigation starts by revisiting the explanatory claims in the original SimSiam. After refuting their claims, we introduce vector decomposition for analyzing the collapse based on the gradient analysis of the l_2-normalized representation vector. This yields a unified perspective on how negative samples and SimSiam alleviate collapse. Such a unified perspective comes timely for understanding the recent progress in SSL.

Attention-based Interpretability with Concept Transformers

- Mattia Rigotti, Christoph Miksovic, Ioana Giurgiu, Thomas Gschwind, Paolo Scotton
- abstract@open-review(Poster): Attention is a mechanism that has been instrumental in driving remarkable performance gains of deep neural network models in a host of visual, NLP and multimodal tasks. One additional notable aspect of attention is that it conveniently exposes the ``reasoning'' behind each particular output generated by the model. Specifically, attention scores over input regions or intermediate features have been interpreted as a measure of the contribution of the attended element to the model inference. While the debate in regard to the interpretability of attention is still not settled, researchers have pointed out the existence of architectures and scenarios that afford a meaningful interpretation of the attention mechanism.

Here we propose the generalization of attention from low-level input features to high-level concepts as a mechanism to ensure the interpretability of attention scores within a given application domain. In particular, we design the ConceptTransformer, a deep learning module that exposes explanations of the output of a model in which it is embedded in terms of attention over user-defined high-level concepts. Such explanations are \emph{plausible} (i.e.\ convincing to the human user) and \emph{faithful} (i.e.\ truly reflective of the reasoning process of the model). Plausibility of such explanations is obtained by construction by training the attention heads to conform with known relations between inputs, concepts and outputs dictated by domain knowledge. Faithfulness is achieved by design by enforcing a linear relation between the transformer value vectors that represent the concepts and their contribution to the classification log-probabilities.

We validate our ConceptTransformer module on established explainability benchmarks and show how it can be used to infuse domain knowledge into classifiers to improve accuracy, and conversely to extract concept-based explanations of classification outputs. Code to reproduce our results is available at: \url{https://github.com/ibm/concept_transformer}.

Inductive Relation Prediction Using Analogy Subgraph Embeddings

- Jiarui Jin, Yangkun Wang, Kounianhua Du, Weinan Zhang, Zheng Zhang, David Wipf, Yong Yu, Quan Gan
- abstract@open-review(Poster): Prevailing methods for relation prediction in heterogeneous graphs aim at learning latent representations (i.e., embeddings) of observed nodes and relations, and thus are limited to the transductive setting where the relation types must be known during training. Here, we propose ANalogy SubGraphEmbeddingLearning (GraphANGEL), a novel relation prediction framework that predicts relations5between each node pair based on the subgraphs containing the pair, as well as other (analogy) subgraphs with the same graph patterns. Each graph pattern explicitly represents a specific logical rule, which contributes to an inductive bias that facilitates generalization to unseen relations and leads to more explainable predictive models. Moreover, our method also removes the limited neighborhood constraint of graph neural networks. Our model consistently outperforms existing models on heterogeneous graph based recommendation as well as knowledge graph completion. We also empirically demonstrate our modelâ€™s capability in generalizing to new relations while producing explainable heat maps of attention scores across the discovered logic.

Reinforcement Learning in Presence of Discrete Markovian Context Evolution

- Hang Ren, Aivar Sootla, Taher Jafferjee, Junxiao Shen, Jun Wang, Haitham Bou Ammar
- abstract@open-review(Poster): We consider a context-dependent Reinforcement Learning (RL) setting, which is characterized by: a) an unknown finite number of not directly observable contexts; b) abrupt (discontinuous) context changes occurring during an episode; and c) Markovian context evolution. We argue that this challenging case is often met in applications and we tackle it using a Bayesian model-based approach and variational inference. We adapt a sticky Hierarchical Dirichlet Process (HDP) prior for model learning, which is arguably best-suited for infinite Markov chain modeling. We then derive a context distillation procedure, which identifies and removes spurious contexts in an unsupervised fashion. We argue that the combination of these two components allows inferring the number of contexts from data thus dealing with the context cardinality assumption. We then find the representation of the optimal policy enabling efficient policy learning using off-the-shelf RL algorithms. Finally, we demonstrate empirically (using gym environments cart-pole swing-up, drone, intersection) that our approach succeeds where state-of-the-art methods of other frameworks fail and elaborate on the reasons for such failures.

Optimal Transport for Long-Tailed Recognition with Learnable Cost Matrix

- Hanyu Peng, Mingming Sun, Ping Li
- abstract@open-review(Poster): It is attracting attention to the long-tailed recognition problem, a burning issue that has become very popular recently. Distinctive from conventional recognition is that it posits that the allocation of the training set is supremely distorted. Predictably, it will pose challenges to the generalisation behaviour of the model. Approaches to these challenges revolve into two groups: firstly, training-aware methods, with the aim of enhancing the generalisability of the model by exploiting its potential in the training period; and secondly, post-hoc correction, liberally coupled with training-aware methods, which is intended to refine the predictions to the extent possible in the post-processing stage, offering the advantages of simplicity and effectiveness. This paper introduces an alternative direction to do the post-hoc correction, which goes beyond the statistical methods. Mathematically, we approach this issue from the perspective of optimal transport (OT), yet, choosing the exact cost matrix when applying OT is challenging and requires expert knowledge of various tasks. To overcome this limitation, we propose to employ linear mapping to learn the cost matrix without necessary configurations adaptively. Testing our methods in practice, along with high efficiency and excellent performance, our method surpasses all previous methods and has the best performance to date.

PriorGrad: Improving Conditional Denoising Diffusion Models with Data-Dependent Adaptive Prior

- Sang-gil Lee, Heeseung Kim, Chaehun Shin, Xu Tan, Chang Liu, Qi Meng, Tao Qin, Wei Chen, Sungroh Yoon, Tie-Yan Liu
- abstract@open-review(Poster): Denoising diffusion probabilistic models have been recently proposed to generate high-quality samples by estimating the gradient of the data density. The framework assumes the prior noise as a standard Gaussian distribution, whereas the corresponding data distribution may be more complicated than the standard Gaussian distribution, which potentially introduces inefficiency in denoising the prior noise into the data sample because of the discrepancy between the data and the prior. In this paper, we propose PriorGrad to improve the efficiency of the conditional diffusion model

(for example, a vocoder using a mel-spectrogram as the condition) by applying an adaptive prior derived from the data statistics based on the conditional information. We formulate the training and sampling procedures of PriorGrad and demonstrate the advantages of an adaptive prior through a theoretical analysis. Focusing on the audio domain, we consider the recently proposed diffusion-based audio generative models based on both the spectral and time domains and show that PriorGrad achieves faster convergence and superior performance, leading to an improved perceptual quality and tolerance to a smaller network capacity, and thereby demonstrating the efficiency of a data-dependent adaptive prior.

Target-Side Input Augmentation for Sequence to Sequence Generation

- Shufang Xie, Ang Lv, Yingce Xia, Lijun Wu, Tao Qin, Tie-Yan Liu, Rui Yan
- abstract@open-review(Poster): Autoregressive sequence generation, a prevalent task in machine learning and natural language processing, generates every target token conditioned on both a source input and previously generated target tokens. Previous data augmentation methods, which have been shown to be effective for the task, mainly enhance source inputs (e.g., injecting noise into the source sequence by random swapping or masking, back translation, etc.) while overlooking the target-side augmentation. In this work, we propose a target-side augmentation method for sequence generation. In training, we use the decoder output probability distributions as soft indicators, which are multiplied with target token embeddings, to build pseudo tokens. These soft pseudo tokens are then used as target tokens to enhance the training. We conduct comprehensive experiments on various sequence generation tasks, including dialog generation, machine translation, and abstractive summarization. Without using any extra labeled data or introducing additional model parameters, our method significantly outperforms strong baselines. The code is available at https://github.com/TARGET-SIDE-DATA-AUG/TSDASG.

UniFormer: Unified Transformer for Efficient Spatial-Temporal Representation Learning

- Kunchang Li, Yali Wang, Gao Peng, Guanglu Song, Yu Liu, Hongsheng Li, Yu Qiao
- abstract@open-review(Poster): It is a challenging task to learn rich and multi-scale spatiotemporal semantics from high-dimensional videos, due to large local redundancy and complex global dependency between video frames. The recent advances in this research have been mainly driven by 3D convolutional neural networks and vision transformers. Although 3D convolution can efficiently aggregate local context to suppress local redundancy from a small 3D neighborhood, it lacks the capability to capture global dependency because of the limited receptive field. Alternatively, vision transformers can effectively capture long-range dependency by self-attention mechanism, while having the limitation on reducing local redundancy with blind similarity comparison among all the tokens in each layer. Based on these observations, we propose a novel Unified transFormer (UniFormer) which seamlessly integrates merits of 3D convolution and spatiotemporal self-attention in a concise transformer format, and achieves a preferable balance between computation and accuracy. Different from traditional transformers, our relation aggregator can tackle both spatiotemporal redundancy and dependency, by learning local and global token affinity respectively in shallow and deep layers. We conduct extensive experiments on the popular video benchmarks, e.g., Kinetics-400, Kinetics-600, and Something-Something V1&V2. With only ImageNet-1K pretraining, our UniFormer achieves 82.9%/84.8% top-1 accuracy on Kinetics-400/Kinetics-600, while requiring 10x fewer GFLOPs than other state-of-the-art methods. For Something-Something V1 and V2, our UniFormer achieves new state-of-the-art performances of 60.9% and 71.2% top-1 accuracy respectively. Code is available at https://github.com/Sense-X/UniFormer.

Inverse Online Learning: Understanding Non-Stationary and Reactionary Policies

- Alex Chan, Alicia Curth, Mihaela van der Schaar
- abstract@open-review(Poster): Human decision making is well known to be imperfect and the ability to analyse such processes individually is crucial when attempting to aid or improve a decision-maker's ability to perform a task, e.g. to alert them to potential biases or oversights on their part. To do so, it is necessary to develop interpretable representations of how agents make decisions and how this process changes over time as the agent learns online in reaction to the accrued experience. To then understand the decision-making processes underlying a set of observed trajectories, we cast the policy inference problem as the inverse to this online learning problem. By interpreting actions within a potential outcomes framework, we introduce a meaningful mapping based on agents choosing an action they believe to have the greatest treatment effect. We introduce a practical algorithm for retrospectively estimating such perceived effects, alongside the process through which agents update them, using a novel architecture built upon an expressive family of deep state-space models. Through application to the analysis of UNOS organ donation acceptance decisions, we demonstrate that our approach can bring valuable insights into the factors that govern decision processes and how they change over time.

Multi-Mode Deep Matrix and Tensor Factorization

- Jicong Fan
- abstract@open-review(Poster): Recently, deep linear and nonlinear matrix factorizations gain increasing attention in the area of machine learning. Existing deep nonlinear matrix factorization methods can only exploit partial nonlinearity of the data and are not effective in handling matrices of which the number of rows is comparable to the number of columns. On the other hand, there is still a gap between deep learning and tensor decomposition. This paper presents a framework of multi-mode deep matrix and tensor factorizations to explore and exploit the full nonlinearity of the data in matrices and tensors. We use the factorization methods to solve matrix and tensor completion problems and prove that our methods have tighter generalization error bounds than conventional matrix and tensor factorization methods. The experiments on synthetic data and real datasets showed that the proposed methods have much higher recovery accuracy than many baselines.

LORD: Lower-Dimensional Embedding of Log-Signature in Neural Rough Differential Equations

- JAEHOON LEE, Jinsung Jeon, Sheo yon Jhin, Jihyeon Hyeong, Jayoung Kim, Minju Jo, Kook Seungji, Noseong Park
- abstract@open-review(Poster): The problem of processing very long time-series data (e.g., a length of more than 10,000) is a long-standing research problem in machine learning. Recently, one breakthrough, called neural rough differential equations (NRDEs), has been proposed and has shown that it is able to process such data. Their main concept is to use the log-signature transform, which is known to be more efficient than the Fourier transform for irregular long time-series, to convert a very long time-series sample into a relatively shorter series of feature vectors. However, the log-signature transform causes non-trivial spatial overheads. To this end, we present the method of LOweR-Dimensional embedding of log-signature (LORD), where we define an NRDE-based autoencoder to implant the higher-depth log-signature knowledge into the lower-depth log-signature. We show that the encoder successfully combines the higher-depth and the lower-depth log-signature knowledge, which greatly stabilizes the training process and increases the model accuracy. In our experiments with benchmark datasets, the improvement ratio by our method is up to 75\% in terms of various classification and forecasting evaluation metrics.

Generalized Natural Gradient Flows in Hidden Convex-Concave Games and GANs

- Andjela Mladenovic, Iosif Sakos, Gauthier Gidel, Georgios Piliouras
- abstract@open-review(Poster): Game-theoretic formulations in machine learning have recently risen in prominence, whereby entire modeling paradigms are best captured as zero-sum games. Despite their popularity, however, their dynamics are still poorly understood. This lack of theory is often substantiated with painful empirical observations of volatile training dynamics and even divergence. Such results highlight the need to develop an appropriate theory with convergence guarantees that are powerful enough to inform practice. This paper studies the generalized Gradient Descent-Ascent (GDA) flow in a large class of non-convex non-concave Zero-Sum games dubbed Hidden Convex-Concave games, a class of games that includes GANs. We focus on two specific geometries: a novel geometry induced by the hidden convex-concave structure that we call the hidden mapping geometry and the

Fisher information geometry. For the hidden mapping geometry, we prove global convergence under mild assumptions. In the case of Fisher information geometry, we provide a complete picture of the dynamics in an interesting special setting of team competition via invariant function analysis.

Offline Neural Contextual Bandits: Pessimism, Optimization and Generalization

- Thanh Nguyen-Tang, Sunil Gupta, A. Tuan Nguyen, Svetha Venkatesh
- abstract@open-review(Poster): Offline policy learning (OPL) leverages existing data collected a priori for policy optimization without any active exploration. Despite the prevalence and recent interest in this problem, its theoretical and algorithmic foundations in function approximation settings remain under-developed. In this paper, we consider this problem on the axes of distributional shift, optimization, and generalization in offline contextual bandits with neural networks. In particular, we propose a provably efficient offline contextual bandit with neural network function approximation that does not require any functional assumption on the reward. We show that our method provably generalizes over unseen contexts under a milder condition for distributional shift than the existing OPL works. Notably, unlike any other OPL method, our method learns from the offline data in an online manner using stochastic gradient descent, allowing us to leverage the benefits of online learning into an offline setting. Moreover, we show that our method is more computationally efficient and has a better dependence on the effective dimension of the neural network than an online counterpart. Finally, we demonstrate the empirical effectiveness of our method in a range of synthetic and real-world OPL problems.

THOMAS: Trajectory Heatmap Output with learned Multi-Agent Sampling

- Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, Fabien Moutarde
- abstract@open-review(Poster): In this paper, we propose THOMAS, a joint multi-agent trajectory prediction framework allowing for an efficient and consistent prediction of multi-agent multi-modal trajectories. We present a unified model architecture for simultaneous agent future heatmap estimation, in which we leverage hierarchical and sparse image generation for fast and memory-efficient inference. We propose a learnable trajectory recombination model that takes as input a set of predicted trajectories for each agent and outputs its consistent reordered recombination. This recombination module is able to realign the initially independent modalities so that they do no collide and are coherent with each other. We report our results on the Interaction multi-agent prediction challenge and rank 1^{st} on the online test leaderboard.

CLEVA-Compass: A Continual Learning Evaluation Assessment Compass to Promote Research Transparency and Comparability

- Martin Mundt, Steven Lang, Quentin Delfosse, Kristian Kersting
- abstract@open-review(Poster): What is the state of the art in continual machine learning? Although a natural question for predominant static benchmarks, the notion to train systems in a lifelong manner entails a plethora of additional challenges with respect to set-up and evaluation. The latter have recently sparked a growing amount of critiques on prominent algorithm-centric perspectives and evaluation protocols being too narrow, resulting in several attempts at constructing guidelines in favor of specific desiderata or arguing against the validity of prevalent assumptions. In this work, we depart from this mindset and argue that the goal of a precise formulation of desiderata is an ill-posed one, as diverse applications may always warrant distinct scenarios. Instead, we introduce the Continual Learning EValuation Assessment Compass: the CLEVA-Compass. The compass provides the visual means to both identify how approaches are practically reported and how works can simultaneously be contextualized in the broader literature landscape. In addition to promoting compact specification in the spirit of recent replication trends, it thus provides an intuitive chart to understand the priorities of individual systems, where they resemble each other, and what elements are missing towards a fair comparison.

Neural Stochastic Dual Dynamic Programming

- Hanjun Dai, Yuan Xue, Zia Syed, Dale Schuurmans, Bo Dai
- abstract@open-review(Poster): Stochastic dual dynamic programming (SDDP) is a state-of-the-art method for solving multi-stage stochastic optimization, widely used for modeling real-world process optimization tasks. Unfortunately, SDDP has a worst-case complexity that scales exponentially in the number of decision variables, which severely limits applicability to only low dimensional problems. To overcome this limitation, we extend SDDP by introducing a trainable neural model that learns to map problem instances to a piece-wise linear value function within intrinsic low-dimension space, which is architected specifically to interact with a base SDDP solver, so that can accelerate optimization performance on new instances. The proposed Neural Stochastic Dual Dynamic Programming ($$\nu$$-SDDP) continually self-improves by solving successive problems. An empirical investigation demonstrates that $$\nu$$-SDDP can significantly reduce problem solving cost without sacrificing solution quality over competitors such as SDDP and reinforcement learning algorithms, across a range of synthetic and real-world process optimization problems.

DemoDICE: Offline Imitation Learning with Supplementary Imperfect Demonstrations

- Geon-Hyeong Kim, Seokin Seo, Jongmin Lee, Wonseok Jeon, HyeongJoo Hwang, Hongseok Yang, Kee-Eung Kim
- abstract@open-review(Poster): We consider offline imitation learning (IL), which aims to mimic the expert's behavior from its demonstration without further interaction with the environment. One of the main challenges in offline IL is to deal with the narrow support of the data distribution exhibited by the expert demonstrations that cover only a small fraction of the state and the action spaces. As a result, offline IL algorithms that rely only on expert demonstrations are very unstable since the situation easily deviates from those in the expert demonstrations. In this paper, we assume additional demonstration data of unknown degrees of optimality, which we call imperfect demonstrations. Under this setting, we propose DemoDICE, which effectively utilizes imperfect demonstrations by matching the stationary distribution of a policy with experts' distribution while penalizing its deviation from the overall demonstrations. Compared with the recent IL algorithms that adopt adversarial minimax training objectives, we substantially stabilize overall learning process by reducing minimax optimization to a direct convex optimization in a principled manner. Using extensive tasks, we show that DemoDICE achieves promising results in the offline IL from expert and imperfect demonstrations.

Learning to Extend Molecular Scaffolds with Structural Motifs

- Krzysztof Maziarz, Henry Richard Jackson-Flux, Pashmina Cameron, Finton Sirockin, Nadine Schneider, Nikolaus Stiefl, Marwin Segler, Marc Brockschmidt
- abstract@open-review(Poster): Recent advancements in deep learning-based modeling of molecules promise to accelerate in silico drug discovery. A plethora of generative models is available, building molecules either atom-by-atom and bond-by-bond or fragment-by-fragment. However, many drug discovery projects require a fixed scaffold to be present in the generated molecule, and incorporating that constraint has only recently been explored. Here, we propose MoLeR, a graph-based model that naturally supports scaffolds as initial seed of the generative procedure, which is possible because it is not conditioned on the generation history. Our experiments show that MoLeR performs comparably to state-of-the-art methods on unconstrained molecular optimization tasks, and outperforms them on scaffold-based tasks, while being an order of magnitude faster to train and sample from than existing approaches. Furthermore, we show the influence of a number of seemingly minor design choices on the overall performance.

Discrepancy-Based Active Learning for Domain Adaptation

- Antoine de Mathelin, FranÃ§ois Deheeger, Mathilde MOUGEOT, Nicolas Vayatis

- abstract@open-review(Poster): The goal of the paper is to design active learning strategies which lead to domain adaptation under an assumption of Lipschitz functions. Building on previous work by Mansour et al. (2009) we adapt the concept of discrepancy distance between source and target distributions to restrict the maximization over the hypothesis class to a localized class of functions which are performing accurate labeling on the source domain. We derive generalization error bounds for such active learning strategies in terms of Rademacher average and localized discrepancy for general loss functions which satisfy a regularity condition. A practical K-medoids algorithm that can address the case of large data set is inferred from the theoretical bounds. Our numerical experiments show that the proposed algorithm is competitive against other state-of-the-art active learning techniques in the context of domain adaptation, in particular on large data sets of around one hundred thousand images.

Gradient Matching for Domain Generalization

- Yuge Shi, Jeffrey Seely, Philip Torr, Siddharth N, Awni Hannun, Nicolas Usunier, Gabriel Synnaeve
- abstract@open-review(Poster): Machine learning systems typically assume that the distributions of training and test sets match closely. However, a critical requirement of such systems in the real world is their ability to generalize to unseen domains. Here, we propose an *inter-domain gradient matching* objective that targets domain generalization by maximizing the inner product between gradients from different domains. Since direct optimization of the gradient inner product can be computationally prohibitive --- it requires computation of second-order derivatives â€“-- we derive a simpler first-order algorithm named Fish that approximates its optimization. We perform experiments on the Wilds benchmark, which captures distribution shift in the real world, as well as the DomainBed benchmark that focuses more on synthetic-to-real transfer. Our method produces competitive results on both benchmarks, demonstrating its effectiveness across a wide range of domain generalization tasks.

Objects in Semantic Topology

- Shuo Yang, Peize Sun, Yi Jiang, Xiaobo Xia, Ruiheng Zhang, Zehuan Yuan, Changhu Wang, Ping Luo, Min Xu
- abstract@open-review(Poster): A more realistic object detection paradigm, Open-World Object Detection, has arised increasing research interests in the community recently. A qualified open-world object detector can not only identify objects of known categories, but also discover unknown objects, and incrementally learn to categorize them when their annotations progressively arrive. Previous works rely on independent modules to recognize unknown categories and perform incremental learning, respectively. In this paper, we provide a unified perspective: Semantic Topology. During the life-long learning of an open-world object detector, all object instances from the same category are assigned to their corresponding pre-defined node in the semantic topology, including the `unknown' category. This constraint builds up discriminative feature representations and consistent relationships among objects, thus enabling the detector to distinguish unknown objects out of the known categories, as well as making learned features of known objects undistorted when learning new categories incrementally. Extensive experiments demonstrate that semantic topology, either randomly-generated or derived from a well-trained language model, could outperform the current state-of-the-art open-world object detectors by a large margin, e.g., the absolute open-set error (the number of unknown instances that are wrongly labeled as known) is reduced from 7832 to 2546, exhibiting the inherent superiority of semantic topology on open-world object detection.

Hidden Parameter Recurrent State Space Models For Changing Dynamics Scenarios

- Vaisakh Shaj, Dieter BÃ¼chler, Rohit Sonker, Philipp Becker, Gerhard Neumann
- abstract@open-review(Poster): Recurrent State-space models (RSSMs) are highly expressive models for learning patterns in time series data and for system identification. However, these models are often based on the assumption that the dynamics are fixed and unchanging, which is rarely the case in real-world scenarios. Many control applications often exhibit tasks with similar, but not identical dynamics, that can be modelled as having a common latent structure. We introduce the Hidden Parameter Recurrent State Space Models (HiP-RSSMs), a framework that parametrizes a family of related state-space models with a low-dimensional set of latent factors. We present a simple and effective way of performing learning and inference over this Gaussian graphical model that avoids approximations like variational inference. We show that HiP-RSSMs outperforms RSSMs and competing multi-task models on several challenging robotic benchmarks both on real systems and simulations.

Graph Neural Network Guided Local Search for the Traveling Salesperson Problem

- Benjamin Hudson, Qingbiao Li, Matthew Malencia, Amanda Prorok
- abstract@open-review(Poster): Solutions to the Traveling Salesperson Problem (TSP) have practical applications to processes in transportation, logistics, and automation, yet must be computed with minimal delay to satisfy the real-time nature of the underlying tasks. However, solving large TSP instances quickly without sacrificing solution quality remains challenging for current approximate algorithms. To close this gap, we present a hybrid data-driven approach for solving the TSP based on Graph Neural Networks (GNNs) and Guided Local Search (GLS). Our model predicts the regret of including each edge of the problem graph in the solution; GLS uses these predictions in conjunction with the original problem graph to find solutions. Our experiments demonstrate that this approach converges to optimal solutions at a faster rate than three recent learning based approaches for the TSP. Notably, we reduce the mean optimality gap on the 100-node problem set from 1.534% to 0.705%, a 2x improvement. When generalizing from 20-node instances to the 100-node problem set, we reduce the optimality gap from 18.845% to 2.622%, a 7x improvement.

On the Pitfalls of Heteroscedastic Uncertainty Estimation with Probabilistic Neural Networks

- Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, Georg Martius
- abstract@open-review(Poster): Capturing aleatoric uncertainty is a critical part of many machine learning systems. In deep learning, a common approach to this end is to train a neural network to estimate the parameters of a heteroscedastic Gaussian distribution by maximizing the logarithm of the likelihood function under the observed data. In this work, we examine this approach and identify potential hazards associated with the use of log-likelihood in conjunction with gradient-based optimizers. First, we present a synthetic example illustrating how this approach can lead to very poor but stable parameter estimates. Second, we identify the culprit to be the log-likelihood loss, along with certain conditions that exacerbate the issue. Third, we present an alternative formulation, termed β-NLL, in which each data point's contribution to the loss is weighted by the β-exponentiated variance estimate. We show that using an appropriate β largely mitigates the issue in our illustrative example. Fourth, we evaluate this approach on a range of domains and tasks and show that it achieves considerable improvements and performs more robustly concerning hyperparameters, both in predictive RMSE and log-likelihood criteria.

Label-Efficient Semantic Segmentation with Diffusion Models

- Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, Artem Babenko
- abstract@open-review(Poster): Denoising diffusion probabilistic models have recently received much research attention since they outperform alternative approaches, such as GANs, and currently provide state-of-the-art generative performance. The superior performance of diffusion models has made them an appealing tool in several applications, including inpainting, super-resolution, and semantic editing. In this paper, we demonstrate that diffusion models can also serve as an instrument for semantic segmentation, especially in the setup when labeled data is scarce. In particular, for several pretrained diffusion models, we investigate the intermediate activations from the networks that perform the Markov step of the reverse diffusion process. We show that these activations effectively capture the semantic information from an input image and appear to be excellent pixel-level representations for the segmentation problem. Based on these observations, we describe a simple segmentation method, which can work even if only a few training images are provided. Our approach significantly outperforms the existing alternatives on several datasets for the same amount of human supervision.

Language model compression with weighted low-rank factorization

- Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, Hongxia Jin
- abstract@open-review(Poster): Factorizing a large matrix into small matrices is a popular strategy for model compression. Singular value decomposition (SVD) plays a vital role in this compression strategy, approximating a learned matrix with fewer parameters. However, SVD minimizes the squared error toward reconstructing the original matrix without gauging the importance of the parameters, potentially giving a larger reconstruction error for those who affect the task accuracy more. In other words, the optimization objective of SVD is not aligned with the trained model's task accuracy. We analyze this previously unexplored problem, make observations, and address it by introducing Fisher information to weigh the importance of parameters affecting the model prediction. This idea leads to our method: Fisher-Weighted SVD (FWSVD). Although the factorized matrices from our approach do not result in smaller reconstruction errors, we find that our resulting task accuracy is much closer to the original model's performance. We perform analysis with the transformer-based language models, showing our weighted SVD largely alleviates the mismatched optimization objectives and can maintain model performance with a higher compression rate. Our method can directly compress a task-specific model while achieving better performance than other compact model strategies requiring expensive model pre-training. Moreover, the evaluation of compressing an already compact model shows our method can further reduce 9% to 30% parameters with an insignificant impact on task accuracy.

Pareto Set Learning for Neural Multi-Objective Combinatorial Optimization

- Xi Lin, Zhiyuan Yang, Qingfu Zhang
- abstract@open-review(Poster): Multiobjective combinatorial optimization (MOCO) problems can be found in many real-world applications. However, exactly solving these problems would be very challenging, particularly when they are NP-hard. Many handcrafted heuristic methods have been proposed to tackle different MOCO problems over the past decades. In this work, we generalize the idea of neural combinatorial optimization, and develop a learning-based approach to approximate the whole Pareto set for a given MOCO problem without further search procedure. We propose a single preference-conditioned model to directly generate approximate Pareto solutions for any trade-off preference, and design an efficient multiobjective reinforcement learning algorithm to train this model. Our proposed method can be treated as a learning-based extension for the widely-used decomposition-based multiobjective evolutionary algorithm (MOEA/D). It uses a single model to accommodate all the possible preferences, whereas other methods use a finite number of solutions to approximate the Pareto set. Experimental results show that our proposed method significantly outperforms some other methods on the multiobjective traveling salesman problem, multiobjective vehicle routing problem, and multiobjective knapsack problem in terms of solution quality, speed, and model efficiency.

Prototypical Contrastive Predictive Coding

- Kyungmin Lee
- abstract@open-review(Poster): Transferring representational knowledge of a model to another is a wide-ranging topic in machine learning. Those applications include the distillation of a large supervised or self-supervised teacher model to a smaller student model or self-supervised learning via self-distillation. Knowledge distillation is an original method to solve these problems, which minimizes a cross-entropy loss between the prototypical probabilistic outputs of teacher and student networks. On the other hand, contrastive learning has shown its competency in transferring representations as they allow students to capture the information of teacher representations. In this paper, we amalgamate the advantages of knowledge distillation and contrastive learning by modeling the critic of a contrastive objective by the prototypical probabilistic discrepancy between two features. We refer to it as prototypical contrastive predictive coding and present efficient implementation using the proposed objective for three distillation tasks: supervised model compression, self-supervised model compression, and self-supervised learning via self-distillation. Through extensive experiments, we validate the effectiveness of our method and show that our method achieves state-of-the-art performance in supervised / self-supervised model compression.

Adversarial Robustness Through the Lens of Causality

- Yonggang Zhang, Mingming Gong, Tongliang Liu, Gang Niu, Xinmei Tian, Bo Han, Bernhard SchÃ¶lkopf, Kun Zhang
- abstract@open-review(Poster): The adversarial vulnerability of deep neural networks has attracted signiï¬ cant attention in machine learning. As causal reasoning has an instinct for modeling distribution change, it is essential to incorporate causality into analyzing this specific type of distribution change induced by adversarial attacks. However, causal formulations of the intuition of adversarial attacks and the development of robust DNNs are still lacking in the literature. To bridge this gap, we construct a causal graph to model the generation process of adversarial examples and define the adversarial distribution to formalize the intuition of adversarial attacks. From the causal perspective, we study the distinction between the natural and adversarial distribution and conclude that the origin of adversarial vulnerability is the focus of models on spurious correlations. Inspired by the causal understanding, we propose the \emph{Causal}-inspired \emph{Adv}ersarial distribution alignment method, CausalAdv, to eliminate the difference between natural and adversarial distributions by considering spurious correlations. Extensive experiments demonstrate the efficacy of the proposed method. Our work is the first attempt towards using causality to understand and mitigate the adversarial vulnerability.

Distributionally Robust Fair Principal Components via Geodesic Descents

- Hieu Vu, Toan Tran, Man-Chung Yue, Viet Anh Nguyen
- abstract@open-review(Poster): Principal component analysis is a simple yet useful dimensionality reduction technique in modern machine learning pipelines. In consequential domains such as college admission, healthcare and credit approval, it is imperative to take into account emerging criteria such as the fairness and the robustness of the learned projection. In this paper, we propose a distributionally robust optimization problem for principal component analysis which internalizes a fairness criterion in the objective function. The learned projection thus balances the trade-off between the total reconstruction error and the reconstruction error gap between subgroups, taken in the min-max sense over all distributions in a moment-based ambiguity set. The resulting optimization problem over the Stiefel manifold can be efficiently solved by a Riemannian subgradient descent algorithm with a sub-linear convergence rate. Our experimental results on real-world datasets show the merits of our proposed method over state-of-the-art baselines.

Understanding and Improving Graph Injection Attack by Promoting Unnoticeability

- Yongqiang Chen, Han Yang, Yonggang Zhang, MA KAILI, Tongliang Liu, Bo Han, James Cheng
- abstract@open-review(Poster): Recently Graph Injection Attack (GIA) emerges as a practical attack scenario on Graph Neural Networks (GNNs), where the adversary can merely inject few malicious nodes instead of modifying existing nodes or edges, i.e., Graph Modification Attack (GMA). Although GIA has achieved promising results, little is known about why it is successful and whether there is any pitfall behind the success. To understand the power of GIA, we compare it with GMA and find that GIA can be provably more harmful than GMA due to its relatively high flexibility. However, the high flexibility will also lead to great damage to the homophily distribution of the original graph, i.e., similarity among neighbors. Consequently, the threats of GIA can be easily alleviated or even prevented by homophily-based defenses designed to recover the original homophily. To mitigate the issue, we introduce a novel constraint â€“ homophily unnoticeability that enforces GIA to preserve the homophily, and propose Harmonious Adversarial Objective (HAO) to instantiate it. Extensive experiments verify that GIA with HAO can break homophily-based defenses and outperform previous GIA attacks by a significant margin. We believe our methods can serve for a more reliable evaluation of the robustness of GNNs.

Learning to Guide and to be Guided in the Architect-Builder Problem

- Paul Barde, Tristan Karch, Derek Nowrouzezahrai, ClÃ©ment Moulin-Frier, Christopher Pal, Pierre-Yves Oudeyer
- abstract@open-review(Poster): We are interested in interactive agents that learn to coordinate, namely, a $builder$ -- which performs actions but ignores the goal of the task, i.e. has no access to rewards -- and an $architect$ which guides the builder towards the goal of the task. We define and explore a formal setting where artificial agents are equipped with mechanisms that allow them to simultaneously learn a task while at the same time evolving a shared communication protocol.
  Ideally, such learning should only rely on high-level communication priors and be able to handle a large variety of tasks and meanings while deriving communication protocols that can be reused across tasks. The field of Experimental Semiotics has shown the extent of human proficiency at learning from a priori unknown instructions meanings. Therefore, we take inspiration from it and present the Architect-Builder Problem (ABP): an asymmetrical setting in which an architect must learn to guide a builder towards constructing a specific structure. The architect knows the target structure but cannot act in the environment and can only send arbitrary messages to the builder. The builder on the other hand can act in the environment, but receives no rewards nor has any knowledge about the task, and must learn to solve it relying only on the messages sent by the architect. Crucially, the meaning of messages is initially not defined nor shared between the agents but must be negotiated throughout learning. Under these constraints, we propose Architect-Builder Iterated Guiding (ABIG), a solution to the Architect-Builder Problem where the architect leverages a learned model of the builder to guide it while the builder uses self-imitation learning to reinforce its guided behavior. To palliate to the non-stationarity induced by the two agents concurrently learning, ABIG structures the sequence of interactions between the agents into interaction frames. We analyze the key learning mechanisms of ABIG and test it in a 2-dimensional instantiation of the ABP where tasks involve grasping cubes, placing them at a given location, or building various shapes. In this environment, ABIG results in a low-level, high-frequency, guiding communication protocol that not only enables an architect-builder pair to solve the task at hand, but that can also generalize to unseen tasks.

Phase Collapse in Neural Networks

- Florentin Guth, John Zarka, StÃ©phane Mallat
- abstract@open-review(Poster): Deep convolutional classifiers linearly separate image classes and improve accuracy as depth increases. They progressively reduce the spatial dimension whereas the number of channels grows with depth. Spatial variability is therefore transformed into variability along channels. A fundamental challenge is to understand the role of non-linearities together with convolutional filters in this transformation. ReLUs with biases are often interpreted as thresholding operators that improve discrimination through sparsity. This paper demonstrates that it is a different mechanism called \emph{phase collapse} which eliminates spatial variability while linearly separating classes. We show that collapsing the phases of complex wavelet coefficients is sufficient to reach the classification accuracy of ResNets of similar depths. However, replacing the phase collapses with thresholding operators that enforce sparsity considerably degrades the performance. We explain these numerical results by showing that the iteration of phase collapses progressively improves separation of classes, as opposed to thresholding non-linearities.

SPIRAL: Self-supervised Perturbation-Invariant Representation Learning for Speech Pre-Training

- Wenyong Huang, Zhenhe Zhang, Yu Ting Yeung, Xin Jiang, Qun Liu
- abstract@open-review(Poster): We introduce a new approach for speech pre-training named SPIRAL which works by learning denoising representation of perturbed data in a teacher-student framework. Specifically, given a speech utterance, we first feed the utterance to a teacher network to obtain corresponding representation. Then the same utterance is perturbed and fed to a student network. The student network is trained to output representation resembling that of the teacher. At the same time, the teacher network is updated as moving average of student's weights over training steps. In order to prevent representation collapse, we apply an in-utterance contrastive loss as pre-training objective and impose position randomization on the input to the teacher. SPIRAL achieves competitive or better results compared to state-of-the-art speech pre-training method wav2vec 2.0, with significant reduction of training cost (80% for BASE model, 65% for LARGE model). Furthermore, we address the problem of noise-robustness that is critical to real-world speech applications. We propose multi-condition pre-training by perturbing the student's input with various types of additive noise. We demonstrate that multi-condition pre-trained SPIRAL models are more robust to noisy speech (9.0% - 13.3% relative word error rate reduction on real noisy test data), compared to applying multi-condition training solely in the fine-tuning stage. Source code is available at https://github.com/huawei-noah/Speech-Backbones/tree/main/SPIRAL.

Improving the Accuracy of Learning Example Weights for Imbalance Classification

- Yuqi Liu, Bin Cao, Jing Fan
- abstract@open-review(Poster): To solve the imbalance classification, methods of weighting examples have been proposed. Recent work has studied to assign adaptive weights to training examples through learning mechanisms, that is, the weights, similar to classification models, are regarded as parameters that need to be learned. However, the algorithms in recent work use local information to approximately optimize the weights, which may lead to inaccurate learning of the weights. In this work, we first propose a novel mechanism of learning with a constraint, which can accurately train the weights and model. Then, we propose a combined method of our learning mechanism and the work by Hu et al., which can promote each other to perform better. Our proposed method can be applied to any type of deep network model. Experiments show that compared with the state-of-the-art algorithms, our method has significant improvement in varieties of settings, including text and image classification over different imbalance ratios, binary and multi-class classification.

Generating Videos with Dynamics-aware Implicit Generative Adversarial Networks

- Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, Jinwoo Shin
- abstract@open-review(Poster): In the deep learning era, long video generation of high-quality still remains challenging due to the spatio-temporal complexity and continuity of videos. Existing prior works have attempted to model video distribution by representing videos as 3D grids of RGB values, which impedes the scale of generated videos and neglects continuous dynamics. In this paper, we found that the recent emerging paradigm of implicit neural representations (INRs) that encodes a continuous signal into a parameterized neural network effectively mitigates the issue. By utilizing INRs of video, we propose dynamics-aware implicit generative adversarial network (DIGAN), a novel generative adversarial network for video generation. Specifically, we introduce (a) an INR-based video generator that improves the motion dynamics by manipulating the space and time coordinates differently and (b) a motion discriminator that efficiently identifies the unnatural motions without observing the entire long frame sequences. We demonstrate the superiority of DIGAN under various datasets, along with multiple intriguing properties, e.g., long video synthesis, video extrapolation, and non-autoregressive video generation. For example, DIGAN improves the previous state-of-the-art FVD score on UCF-101 by 30.7% and can be trained on 128 frame videos of 128x128 resolution, 80 frames longer than the 48 frames of the previous state-of-the-art method.

Efficient Learning of Safe Driving Policy via Human-AI Copilot Optimization

- Quanyi Li, Zhenghao Peng, Bolei Zhou
- abstract@open-review(Poster): Human intervention is an effective way to inject human knowledge into the training loop of reinforcement learning, which can bring fast learning and ensured training safety. Given the very limited budget of human intervention, it remains challenging to design when and how human expert interacts with the learning agent in the training. In this work, we develop a novel human-in-the-loop learning method called Human-AI Copilot Optimization (HACO).To allow the agent's sufficient exploration in the risky environments while ensuring the training safety, the human expert can take over the control and demonstrate how to avoid probably dangerous situations or trivial behaviors. The proposed HACO then effectively utilizes the data both from the trial-and-error exploration and human's partial demonstration to train a high-performing agent. HACO extracts proxy state-action values from partial human demonstration and optimizes the agent to improve the proxy values meanwhile reduce the human interventions. The

experiments show that HACO achieves a substantially high sample efficiency in the safe driving benchmark. HACO can train agents to drive in unseen traffic scenarios with a handful of human intervention budget and achieve high safety and generalizability, outperforming both reinforcement learning and imitation learning baselines with a large margin. Code and demo video are included in the supplementary materials.

Enhancing Cross-lingual Transfer by Manifold Mixup

- Huiyun Yang, Huadong Chen, Hao Zhou, Lei Li
- abstract@open-review(Poster): Based on large-scale pre-trained multilingual representations, recent cross-lingual transfer methods have achieved impressive transfer performances. However, the performance of target languages still lags far behind the source language. In this paper, our analyses indicate such a performance gap is strongly associated with the cross-lingual representation discrepancy. To achieve better cross-lingual transfer performance, we propose the cross-lingual manifold mixup (X-Mixup) method, which adaptively calibrates the representation discrepancy and gives a compromised representation for target languages. Experiments on the XTREME benchmark show X-Mixup achieves 1.8% performance gains on multiple text understanding tasks, compared with strong baselines, and significantly reduces the cross-lingual representation discrepancy.

Evolutionary Diversity Optimization with Clustering-based Selection for Reinforcement Learning

- Yutong Wang, Ke Xue, Chao Qian
- abstract@open-review(Poster): Reinforcement Learning (RL) has achieved significant successes, which aims to obtain a single policy maximizing the expected cumulative rewards for a given task. However, in many real-world scenarios, e.g., navigating in complex environments and controlling robots, one may need to find a set of policies having both high rewards and diverse behaviors, which can bring better exploration and robust few-shot adaptation. Recently, some methods have been developed by using evolutionary techniques, including iterative reproduction and selection of policies. However, due to the inefficient selection mechanisms, these methods cannot fully guarantee both high quality and diversity. In this paper, we propose EDO-CS, a new Evolutionary Diversity Optimization algorithm with Clustering-based Selection. In each iteration, the policies are divided into several clusters based on their behaviors, and a high-quality policy is selected from each cluster for reproduction. EDO-CS also adaptively balances the importance between quality and diversity in the reproduction process. Experiments on various (i.e., deceptive and multi-modal) continuous control tasks, show the superior performance of EDO-CS over previous methods, i.e., EDO-CS can achieve a set of policies with both high quality and diversity efficiently while previous methods cannot.

CURVATURE-GUIDED DYNAMIC SCALE NETWORKS FOR MULTI-VIEW STEREO

- Khang Truong Giang, Soohwan Song, Sungho Jo
- abstract@open-review(Poster): Multi-view stereo (MVS) is a crucial task for precise 3D reconstruction. Most recent studies tried to improve the performance of matching cost volume in MVS by introducing a skilled design to cost formulation or cost regularization. In this paper, we focus on learning robust feature extraction to enhance the performance of matching costs, without need of heavy computation in the other steps. In particular, we present a dynamic scale feature extraction network, namely, CDSFNet. It is composed of multiple novel convolution layers, each of which can select a proper patch scale for each pixel guided by the normal curvature of image surface. As a result, CDFSNet can estimate the optimal patch scales to learn discriminative features for accurate matching computation between reference and source images. By combining the robust extracted features with an appropriate cost formulation strategy, our final MVS architecture can estimate depth maps more precisely. Extensive experiments showed that the proposed method outperforms other state-of-the-art methods on complex outdoor scenes. It significantly improves the completeness of reconstructed models. Moreover, the method can process the high resolution with faster run-time and lower memory compared to the other MVS methods.

Near-optimal Offline Reinforcement Learning with Linear Representation: Leveraging Variance Information with Pessimism

- Ming Yin, Yaqi Duan, Mengdi Wang, Yu-Xiang Wang
- abstract@open-review(Poster): Offline reinforcement learning, which seeks to utilize offline/historical data to optimize sequential decision-making strategies, has gained surging prominence in recent studies. Due to the advantage that appropriate function approximators can help mitigate the sample complexity burden in modern reinforcement learning problems, existing endeavors usually enforce powerful function representation models (e.g. neural networks) to learn the optimal policies. However, a precise understanding of the statistical limits with function representations, remains elusive, even when such a representation is linear.

Towards this goal, we study the statistical limits of offline reinforcement learning with linear model representations. To derive the tight offline learning bound, we design the variance-aware pessimistic value iteration (VAPVI), which adopts the conditional variance information of the value function for time-inhomogeneous episodic linear Markov decision processes (MDPs). VAPVI leverages estimated variances of the value functions to reweight the Bellman residuals in the least-square pessimistic value iteration and provides improved offline learning bounds over the best-known existing results (whereas the Bellman residuals are equally weighted by design). More importantly, our learning bounds are expressed in terms of system quantities, which provide natural instance-dependent characterizations that previous results are short of. We hope our results draw a clearer picture of what offline learning should look like when linear representations are provided.

Exploring extreme parameter compression for pre-trained language models

- Benyou Wang, Yuxin Ren, Lifeng Shang, Xin Jiang, Qun Liu
- abstract@open-review(Poster): Recent work explored the potential of large-scale Transformer-based pre-trained models, especially Pre-trained Language Models (PLMs) in natural language processing. This raises many concerns from various perspectives, e.g., financial costs and carbon emissions. Compressing PLMs like BERT with negligible performance loss for faster inference and cheaper deployment has attracted much attention. In this work, we aim to explore larger compression ratios for PLMs, among which tensor decomposition is a potential but under-investigated one. By comparing existing decomposition methods, Tucker decomposition is found to be parameter-efficient for compression. Two decomposition and reconstruction protocols are further proposed to improve the effectiveness and efficiency of Tucker decomposition in parameter compression. Our compressed BERT with ${1}/{7}$ parameters in Transformer layers performs on-par with, sometimes slightly better than the original BERT in GLUE benchmark. A tiny version achieves 96.7\% performance of BERT-base with $ {1}/{48} $ encoder parameters (i.e., less than 2M parameters excluding the embedding layer) and \textbf{$2.7 \times$} faster on inference. To show that the proposed method is orthogonal to existing compression methods like knowledge distillation, we also explore the benefit of the proposed method on a distilled BERT.

Local Feature Swapping for Generalization in Reinforcement Learning

- David Bertoin, Emmanuel Rachelson
- abstract@open-review(Poster): Over the past few years, the acceleration of computing resources and research in Deep Learning has led to significant practical successes in a range of tasks, including in particular in computer vision. Building on these advances, reinforcement learning has also seen a leap forward with the emergence of agents capable of making decisions directly from visual observations. Despite these successes, the over-parametrization of neural architectures leads to memorization of the data used during training and thus to a lack of generalization. Reinforcement learning agents based on visual inputs also suffer from this phenomenon by erroneously correlating rewards with unrelated visual features such as background elements. To alleviate this problem, we introduce a new regularization layer consisting of channel-consistent local permutations (CLOP) of the feature maps. The proposed permutations induce robustness to spatial correlations and help prevent overfitting behaviors in RL. We demonstrate, on the OpenAI Procgen

Benchmark, that RL agents trained with the CLOP layer exhibit robustness to visual changes and better generalization properties than agents trained using other state-of-the-art regularization techniques.

Open-vocabulary Object Detection via Vision and Language Knowledge Distillation

- Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, Yin Cui
- abstract@open-review(Poster): We aim at advancing open-vocabulary object detection, which detects objects described by arbitrary text inputs. The fundamental challenge is the availability of training data. It is costly to further scale up the number of classes contained in existing object detection datasets. To overcome this challenge, we propose ViLD, a training method via Vision and Language knowledge Distillation. Our method distills the knowledge from a pretrained open-vocabulary image classification model (teacher) into a two-stage detector (student). Specifically, we use the teacher model to encode category texts and image regions of object proposals. Then we train a student detector, whose region embeddings of detected boxes are aligned with the text and image embeddings inferred by the teacher. We benchmark on LVIS by holding out all rare categories as novel categories that are not seen during training. ViLD obtains 16.1 mask APr with a ResNet-50 backbone, even outperforming the supervised counterpart by 3.8. When trained with a stronger teacher model ALIGN, ViLD achieves 26.3 APr. The model can directly transfer to other datasets without finetuning, achieving 72.2 AP50 on PASCAL VOC, 36.6 AP on COCO and 11.8 AP on Objects365. On COCO, ViLD outperforms the previous state-of-the-art (Zareian et al., 2021) by 4.8 on novel AP and 11.4 on overall AP. Code and demo are open-sourced at https://github.com/tensorflow/tpu/tree/master/models/official/detection/projects/vild.

Model-Based Offline Meta-Reinforcement Learning with Regularization

- Sen Lin, Jialin Wan, Tengyu Xu, Yingbin Liang, Junshan Zhang
- abstract@open-review(Poster): Existing offline reinforcement learning (RL) methods face a few major challenges, particularly the distributional shift between the learned policy and the behavior policy. Offline Meta-RL is emerging as a promising approach to address these challenges, aiming to learn an informative meta-policy from a collection of tasks. Nevertheless, as shown in our empirical studies, offline Meta-RL could be outperformed by offline single-task RL methods on tasks with good quality of datasets, indicating that a right balance has to be delicately calibrated between "exploring" the out-of-distribution state-actions by following the meta-policy and "exploiting" the offline dataset by staying close to the behavior policy. Motivated by such empirical analysis, we propose model-based offline $\text{\bf Me}$ta-RL with $\text{\bf r}$egularized $\text{\bf P}$olicy $\text{\bf O}$ptimization (MerPO), which learns a meta-model for efficient task structure inference and an informative meta-policy for safe exploration of out-of-distribution state-actions. In particular, we devise a new meta-Regularized model-based Actor-Critic (RAC) method for within-task policy optimization, as a key building block of MerPO, using both conservative policy evaluation and regularized policy improvement; and the intrinsic tradeoff therein is achieved via striking the right balance between two regularizers, one based on the behavior policy and the other on the meta-policy. We theoretically show that the learnt policy offers guaranteed improvement over both the behavior policy and the meta-policy, thus ensuring the performance improvement on new tasks via offline Meta-RL. Our experiments corroborate the superior performance of MerPO over existing offline Meta-RL methods.

Scale Mixtures of Neural Network Gaussian Processes

- Hyungi Lee, Eunggu Yun, Hongseok Yang, Juho Lee
- abstract@open-review(Poster): Recent works have revealed that infinitely-wide feed-forward or recurrent neural networks of any architecture correspond to Gaussian processes referred to as NNGP. While these works have extended the class of neural networks converging to Gaussian processes significantly, however, there has been little focus on broadening the class of stochastic processes that such neural networks converge to. In this work, inspired by the scale mixture of Gaussian random variables, we propose the scale mixture of NNGP for which we introduce a prior distribution on the scale of the last-layer parameters. We show that simply introducing a scale prior on the last-layer parameters can turn infinitely-wide neural networks of any architecture into a richer class of stochastic processes. With certain scale priors, we obtain heavy-tailed stochastic processes, and in the case of inverse gamma priors, we recover Studentâ€™s t processes. We further analyze the distributions of the neural networks initialized with our prior setting and trained with gradient descents and obtain similar results as for NNGP. We present a practical posterior-inference algorithm for the scale mixture of NNGP and empirically demonstrate its usefulness on regression and classification tasks. In particular, we show that in both tasks, the heavy-tailed stochastic processes obtained from our framework are robust to out-of-distribution data.

A Johnson-Lindenstrauss Framework for Randomly Initialized CNNs

- Ido Nachum, Jan Hazla, Michael Gastpar, Anatoly Khina
- abstract@open-review(Poster): How does the geometric representation of a dataset change after the application of each randomly initialized layer of a neural network? The celebrated Johnson-Lindenstrauss lemma answers this question for linear fully-connected neural networks (FNNs), stating that the geometry is essentially preserved. For FNNs with the ReLU activation, the angle between two input contracts according to a known mapping. The question for non-linear convolutional neural networks (CNNs) becomes much more intricate. To answer this question, we introduce a geometric framework. For linear CNNs, we show that the Johnson--Lindenstrauss lemma continues to hold, namely, that the angle between two inputs is preserved. For CNNs with ReLU activation, on the other hand, the behavior is richer: The angle between the outputs contracts, where the level of contraction depends on the nature of the inputs. In particular, after one layer, the geometry of natural images is essentially preserved, whereas for Gaussian correlated inputs, CNNs exhibit the same contracting behavior as FNNs with ReLU activation.

Hindsight: Posterior-guided training of retrievers for improved open-ended generation

- Ashwin Paranjape, Omar Khattab, Christopher Potts, Matei Zaharia, Christopher D Manning
- abstract@open-review(Poster): Many text generation systems benefit from retrieving passages from a textual knowledge corpus (e.g., Wikipedia) and using them to generate the output. For open-ended generation tasks, like generating informative utterances in conversations, many varied passages z are relevant to the context x but few are relevant to the observed next utterance y (label). For such tasks, existing methods (that jointly train the retriever and generator) underperform: during training the top-k context-relevant retrieved passages might not contain the label-relevant passage and the generator may hence not learn a preference to ground its generated output in them. We propose using an additional guide-retriever that also conditions on the observed label y and â€œin hindsightâ€ retrieves label-relevant passages during training. We maximize the evidence lower bound (ELBo) to jointly train the guide-retriever $Q(z|x,y)$ with the standard retriever $P_\eta(z|x)$ and the generator $P_\theta(y|x,z)$ and find that ELBo has better inductive biases than prior work. For informative conversations from the Wizard of Wikipedia dataset, with our posterior-guided training, the retriever finds passages with higher relevance in the top-10 (23% relative improvement), the generatorâ€™s responses are more grounded in the retrieved passage (19% relative improvement) and the end-to-end system produces better overall output (6.4% relative improvement).

Self-Supervised Graph Neural Networks for Improved Electroencephalographic Seizure Analysis

- Siyi Tang, Jared Dunnmon, Khaled Kamal Saab, Xuan Zhang, Qianying Huang, Florian Dubost, Daniel Rubin, Christopher Lee-Messer
- abstract@open-review(Poster): Automated seizure detection and classification from electroencephalography (EEG) can greatly improve seizure diagnosis and treatment. However, several modeling challenges remain unaddressed in prior automated seizure detection and classification studies: (1) representing non-Euclidean data structure in EEGs, (2) accurately classifying rare seizure types, and (3) lacking a quantitative interpretability approach to measure model ability to localize seizures. In this study, we address these challenges by (1) representing the spatiotemporal dependencies in EEGs using a graph neural network (GNN) and proposing two EEG graph structures that capture the electrode geometry or dynamic brain connectivity, (2) proposing a self-

supervised pre-training method that predicts preprocessed signals for the next time period to further improve model performance, particularly on rare seizure types, and (3) proposing a quantitative model interpretability approach to assess a modelâ€™s ability to localize seizures within EEGs. When evaluating our approach on seizure detection and classification on a large public dataset (5,499 EEGs), we find that our GNN with self-supervised pre-training achieves 0.875 Area Under the Receiver Operating Characteristic Curve on seizure detection and 0.749 weighted F1-score on seizure classification, outperforming previous methods for both seizure detection and classification. Moreover, our self-supervised pre-training strategy significantly improves classification of rare seizure types (e.g. 47 points increase in combined tonic seizure accuracy over baselines). Furthermore, quantitative interpretability analysis shows that our GNN with self-supervised pre-training precisely localizes 25.4% focal seizures, a 21.9 point improvement over existing CNNs. Finally, by superimposing the identified seizure locations on both raw EEG signals and EEG graphs, our approach could provide clinicians with an intuitive visualization of localized seizure regions.

Group-based Interleaved Pipeline Parallelism for Large-scale DNN Training

- PengCheng Yang, Xiaoming Zhang, Wenpeng Zhang, Ming Yang, Hong Wei
- abstract@open-review(Poster): The recent trend of using large-scale deep neural networks (DNN) to boost performance has propelled the development of the parallel pipelining technique for efficient DNN training, which has resulted in the development of several prominent pipelines such as GPipe, PipeDream, and PipeDream-2BW. However, the current leading pipeline PipeDream-2BW still suffers from two major drawbacks, i.e., the excessive memory redundancy and the delayed weight updates across all stages. In this work, we propose a novel pipeline named WPipe, which achieves better memory efficiency and fresher weight updates. WPipe uses a novel pipelining scheme that divides model partitions into two groups. It moves the forward pass of the next period of weight updates to the front of the backward pass of the current period of weight updates in the first group, retains the order in the second group, and updates each group alternatively. This scheme can eliminate half of the delayed gradients and memory redundancy compared to PipeDream-2BW. The experiments, which train large BERT language models, show that compared to PipeDream-2BW, WPipe achieves $1.4\times$ acceleration and reduces the memory footprint by 36%, without nearly sacrificing any final model accuracy.

Minimax Optimality (Probably) Doesn't Imply Distribution Learning for GANs

- Sitan Chen, Jerry Li, Yuanzhi Li, Raghu Meka
- abstract@open-review(Poster): Arguably the most fundamental question in the theory of generative adversarial networks (GANs) is to understand when GANs can actually learn the underlying distribution. Theoretical and empirical evidence (see e.g. Arora-Risteski-Zhang '18) suggest local optimality of the empirical training objective is insufficient, yet it does not rule out the possibility that achieving a true population minimax optimal solution might imply distribution learning. In this paper, we show that standard cryptographic assumptions imply that this stronger condition is still insufficient. Namely, we show that if local pseudorandom generators (PRGs) exist, then for a large family of natural target distributions, there are ReLU network generators of constant depth and poly size which take Gaussian random seeds so that (i) the output is far in Wasserstein distance from the target distribution, but (ii) no polynomially large Lipschitz discriminator ReLU network can detect this. This implies that even achieving a population minimax optimal solution to the Wasserstein GAN objective is likely insufficient for distribution learning. Our techniques reveal a deep connection between GANs and PRGs, which we believe will lead to further insights into the computational landscape of GANs.

Offline Reinforcement Learning with Value-based Episodic Memory

- Xiaoteng Ma, Yiqin Yang, Hao Hu, Jun Yang, Chongjie Zhang, Qianchuan Zhao, Bin Liang, Qihan Liu
- abstract@open-review(Poster): Offline reinforcement learning (RL) shows promise of applying RL to real-world problems by effectively utilizing previously collected data. Most existing offline RL algorithms use regularization or constraints to suppress extrapolation error for actions outside the dataset. In this paper, we adopt a different framework, which learns the V-function instead of the Q-function to naturally keep the learning procedure within the support of an offline dataset. To enable effective generalization while maintaining proper conservatism in offline learning, we propose Expectile V-Learning (EVL), which smoothly interpolates between the optimal value learning and behavior cloning. Further, we introduce implicit planning along offline trajectories to enhance learned V-values and accelerate convergence. Together, we present a new offline method called Value-based Episodic Memory (VEM). We provide theoretical analysis for the convergence properties of our proposed VEM method, and empirical results in the D4RL benchmark show that our method achieves superior performance in most tasks, particularly in sparse-reward tasks.

MonoDistill: Learning Spatial Features for Monocular 3D Object Detection

- Zhiyu Chong, Xinzhu Ma, Hong Zhang, Yuxin Yue, Haojie Li, Zhihui Wang, Wanli Ouyang
- abstract@open-review(Poster): 3D object detection is a fundamental and challenging task for 3D scene understanding, and the monocular-based methods can serve as an economical alternative to the stereo-based or LiDAR-based methods. However, accurately locating objects in the 3D space from a single image is extremely difficult due to the lack of spatial cues. To mitigate this issue, we propose a simple and effective scheme to introduce the spatial information from LiDAR signals to the monocular 3D detectors, without introducing any extra cost in the inference phase. In particular, we first project the LiDAR signals into the image plane and align them with the RGB images. After that, we use the resulting data to train a 3D detector (LiDAR Net) using the same architecture as the baseline model. Finally, this LiDAR Net can serve as the teacher to transfer the learned knowledge to the baseline model. Experimental results show that the proposed method can significantly boost the performance of the baseline model and ranks the 1^{st} place among all monocular-based methods on the KITTI benchmark. Besides, extensive ablation studies are conducted, which further prove the effectiveness of each part of our designs and illustrate what the baseline model has learned from the LiDAR Net.

EXACT: Scalable Graph Neural Networks Training via Extreme Activation Compression

- Zirui Liu, Kaixiong Zhou, Fan Yang, Li Li, Rui Chen, Xia Hu
- abstract@open-review(Poster): Training Graph Neural Networks (GNNs) on large graphs is a fundamental challenge due to the high memory usage, which is mainly occupied by activations (e.g., node embeddings). Previous works usually focus on reducing the number of nodes retained in memory. In parallel, unlike what has been developed for other types of neural networks, training with compressed activation maps is less explored for GNNs. This extension is notoriously difficult to implement due to the miss of necessary tools in common graph learning packages. To unleash the potential of this direction, we provide { an} optimized GPU implementation which supports training GNNs with compressed activations. Based on the implementation, we propose a memory-efficient framework called ``EXACT'', which for the first time demonstrate the potential and evaluate the feasibility of training GNNs with compressed activations. We systematically analyze the trade-off among the memory saving, time overhead, and accuracy drop. In practice, EXACT can reduce the memory footprint of activations by up to $32\times$ with 0.2-0.5% accuracy drop and 10-25% time overhead across different models and datasets. We implement EXACT as an extension for Pytorch Geometric and Pytorch. In practice, for Pytorch Geometric, EXACT can trim down the hardware requirement of training a three-layer full-batch GraphSAGE on \textit{ogbn-products} from a 48GB GPU to a 12GB GPU.

Provably convergent quasistatic dynamics for mean-field two-player zero-sum games

- Chao Ma, Lexing Ying
- abstract@open-review(Poster): In this paper, we study the problem of finding mixed Nash equilibrium for mean-field two-player zero-sum games. Solving this problem requires optimizing over two probability distributions. We consider a quasistatic Wasserstein gradient flow dynamics in which one probability distribution follows the Wasserstein gradient flow, while the other one is always at the equilibrium. Theoretical analysis are conducted on this dynamics, showing its convergence to the mixed Nash equilibrium under mild conditions. Inspired by the continuous dynamics of probability

distributions, we derive a quasistatic Langevin gradient descent method with inner-outer iterations, and test the method on different problems, including training mixture of GANs.

W-CTC: a Connectionist Temporal Classification Loss with Wild Cards

- Xingyu Cai, Jiahong Yuan, Yuchen Bian, Guangxu Xun, Jiaji Huang, Kenneth Church
- abstract@open-review(Poster): Connectionist Temporal Classification (CTC) loss is commonly used in sequence learning applications. For example, in Automatic Speech Recognition (ASR) task, the training data consists of pairs of audio (input sequence) and text (output label),without temporal alignment information. Standard CTC computes a loss by aggregating over all possible alignment paths, that map the entire sequence to the entire label (full alignment). However, in practice, there are often cases where the label is incomplete. Specifically, we solve the partial alignment problem where the label only matches a middle part of the sequence. This paper proposes the wild-card CTC (W-CTC) to address this issue, by padding wild-cards at both ends of the labels. Consequently, the proposed W-CTC improves the standard CTC via aggregating over even more alignment paths. Evaluations on a number of tasks in speech and vision domains, show that the proposed W-CTC consistently outperforms the standard CTC by a large margin when label is incomplete. The effectiveness of the proposed method is further confirmed in an ablation study.

Bandit Learning with Joint Effect of Incentivized Sampling, Delayed Sampling Feedback, and Self-Reinforcing User Preferences

- Tianchen Zhou, Jia Liu, Chaosheng Dong, Yi Sun
- abstract@open-review(Poster): In this paper, we consider a new multi-armed bandit (MAB) framework motivated by three common complications in online recommender systems in practice: (i) the platform (learning agent) cannot sample an intended product directly and has to incentivize customers to select this product (e.g., promotions and coupons); (ii) customer feedbacks are often received later than their selection times; and (iii) customer preferences among products are influenced and reinforced by historical feedbacks. From the platform's perspective, the goal of the MAB framework is to maximize total reward without incurring excessive incentive costs. A major challenge of this MAB framework is that the loss of information caused by feedback delay complicates both user preference evolution and arm incentivizing decisions, both of which are already highly non-trivial even by themselves. Toward this end, we first propose a policy called ``UCB-Filtering-with-Delayed-Feedback'' (UCB-FDF) policy for this new MAB framework. In our analysis, we consider delayed feedbacks that can have either arm-independent or arm-dependent distributions. In both cases, we allow unbounded support for the random delays, i.e., the random delay can be infinite. We show that the delay impacts in both cases can still be upper bounded by an additive penalty on both the regret and total incentive costs. This further implies that logarithmic regret and incentive cost growth rates are achievable under this new MAB framework. Experimental results corroborate our theoretical analysis on both regret and incentive costs.

AdaAug: Learning Class- and Instance-adaptive Data Augmentation Policies

- Tsz-Him Cheung, Dit-Yan Yeung
- abstract@open-review(Poster): Data augmentation is an effective way to improve the generalization capability of modern deep learning models. However, the underlying augmentation methods mostly rely on handcrafted operations. Moreover, an augmentation policy useful to one dataset may not transfer well to other datasets. Therefore, Automated Data Augmentation (AutoDA) methods, like \textit{AutoAugment} and \textit{Population-based Augmentation}, have been proposed recently to automate the process of searching for optimal augmentation policies. However, the augmentation policies found are not adaptive to the dataset used, hindering the effectiveness of these AutoDA methods. In this paper, we propose a novel AutoDA method called \texttt{AdaAug} to efficiently learn adaptive augmentation policies in a class-dependent and potentially instance-dependent manner. Our experiments show that the adaptive augmentation policies learned by our method transfer well to unseen datasets such as the Oxford Flowers, Oxford-IIT Pets, FGVC Aircraft, and Stanford Cars datasets when compared with other AutoDA baselines. In addition, our method also achieves state-of-the-art performance on the CIFAR-10, CIFAR-100, and SVHN datasets.

Unsupervised Semantic Segmentation by Distilling Feature Correspondences

- Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, William T. Freeman
- abstract@open-review(Poster): Unsupervised semantic segmentation aims to discover and localize semantically meaningful categories within image corpora without any form of annotation. To solve this task, algorithms must produce features for every pixel that are both semantically meaningful and compact enough to form distinct clusters. Unlike previous works which achieve this with a single end-to-end framework, we propose to separate feature learning from cluster compactification. Empirically, we show that current unsupervised feature learning frameworks already generate dense features whose correlations are semantically consistent. This observation motivates us to design STEGO ($\textbf{S}$elf-supervised $\textbf{T}$ransformer with $\textbf{E}$nergy-based $\textbf{G}$raph $\textbf{O}$ptimization), a novel framework that distills unsupervised features into high-quality discrete semantic labels. At the core of STEGO is a novel contrastive loss function that encourages features to form compact clusters while preserving their association pattern. STEGO yields a significant improvement over the prior state of the art, on both the CocoStuff ($\textbf{+14 mIoU}$) and Cityscapes ($\textbf{+9 mIoU}$) semantic segmentation challenges.

Axiomatic Explanations for Visual Search, Retrieval, and Similarity Learning

- Mark Hamilton, Scott Lundberg, Stephanie Fu, Lei Zhang, William T. Freeman
- abstract@open-review(Poster): Visual search, recommendation, and contrastive similarity learning power technologies that impact billions of users worldwide. Modern model architectures can be complex and difficult to interpret, and there are several competing techniques one can use to explain a search engine's behavior. We show that the theory of fair credit assignment provides a unique axiomatic solution that generalizes several existing recommendation- and metric-explainability techniques in the literature. Using this formalism, we show when existing approaches violate "fairness" and derive methods that sidestep these shortcomings and naturally handle counterfactual information. More specifically, we show existing approaches implicitly approximate second-order Shapley-Taylor indices and extend CAM, GradCAM, LIME, SHAP, SBSM, and other methods to search engines. These extensions can extract pairwise correspondences between images from trained opaque-box models. We also introduce a fast kernel-based method for estimating Shapley-Taylor indices that require orders of magnitude fewer function evaluations to converge. Finally, we show that these game-theoretic measures yield more consistent explanations for image similarity architectures.

Graph-Relational Domain Adaptation

- Zihao Xu, Hao He, Guang-He Lee, Bernie Wang, Hao Wang
- abstract@open-review(Poster): Existing domain adaptation methods tend to treat every domain equally and align them all perfectly. Such uniform alignment ignores topological structures among different domains; therefore it may be beneficial for nearby domains, but not necessarily for distant domains. In this work, we relax such uniform alignment by using a domain graph to encode domain adjacency, e.g., a graph of states in the US with each state as a domain and each edge indicating adjacency, thereby allowing domains to align flexibly based on the graph structure. We generalize the existing adversarial learning framework with a novel graph discriminator using encoding-conditioned graph embeddings. Theoretical analysis shows that at equilibrium, our method recovers classic domain adaptation when the graph is a clique, and achieves non-trivial alignment for other types of graphs. Empirical results show that our approach successfully generalizes uniform alignment, naturally incorporates domain information represented by graphs, and improves upon existing domain adaptation methods on both synthetic and real-world datasets.

Revisit Kernel Pruning with Lottery Regulated Grouped Convolutions

- Shaochen Zhong, Guanqun Zhang, Ningjia Huang, Shuai Xu
- abstract@open-review(Poster): Structured pruning methods which are capable of delivering a densely pruned network are among the most popular techniques in the realm of neural network pruning, where most methods prune the original network at a filter or layer level. Although such methods may provide immediate compression and acceleration benefits, we argue that the blanket removal of an entire filter or layer may result in undesired accuracy loss. In this paper, we revisit the idea of kernel pruning (to only prune one or several $k \times k$ kernels out of a 3D-filter), a heavily overlooked approach under the context of structured pruning. This is because kernel pruning will naturally introduce sparsity to filters within the same convolutional layer â€" thus, making the remaining network no longer dense. We address this problem by proposing a versatile grouped pruning framework where we first cluster filters from each convolutional layer into equal-sized groups, prune the grouped kernels we deem unimportant from each filter group, then permute the remaining filters to form a densely grouped convolutional architecture (which also enables the parallel computing capability) for fine-tuning. Specifically, we consult empirical findings from a series of literature regarding $\textit{Lottery Ticket Hypothesis}$ to determine the optimal clustering scheme per layer, and develop a simple yet cost-efficient greedy approximation algorithm to determine which group kernels to keep within each filter group. Extensive experiments also demonstrate our method often outperforms comparable SOTA methods with lesser data augmentation needed, smaller fine-tuning budget required, and sometimes even much simpler procedure executed (e.g., one-shot v. iterative). Please refer to our GitHub repository (https://github.com/choH/lottery_regulated_grouped_kernel_pruning) for code.

Bi-linear Value Networks for Multi-goal Reinforcement Learning

- Zhang-Wei Hong, Ge Yang, Pulkit Agrawal
- abstract@open-review(Poster): Universal value functions are a core component of off-policy multi-goal reinforcement learning. The de-facto paradigm is to approximate Q(s, a, g) using monolithic neural networks which lack inductive biases to produce complex interactions between the state s and the goal g. In this work, we propose a bilinear decomposition that represents the Q-value via a low-rank approximation in the form of a dot product between two vector fields. The first vector field, f(s, a), captures the environment's local dynamics at the state s; whereas the second component, Ï•(s, g), captures the global relationship between the current state and the goal. We show that our bilinear decomposition scheme improves sample efficiency over the original monolithic value approximators, and transfer better to unseen goals. We demonstrate significant learning speed-up over a variety of tasks on a simulated robot arm, and the challenging task of dexterous manipulation with a Shadow hand.

No One Representation to Rule Them All: Overlapping Features of Training Methods

- Raphael Gontijo-Lopes, Yann Dauphin, Ekin Dogus Cubuk
- abstract@open-review(Poster): Despite being able to capture a range of features of the data, high accuracy models trained with supervision tend to make similar predictions. This seemingly implies that high-performing models share similar biases regardless of training methodology, which would limit ensembling benefits and render low-accuracy models as having little practical use. Against this backdrop, recent work has developed quite different training techniques, such as large-scale contrastive learning, yielding competitively high accuracy on generalization and robustness benchmarks. This motivates us to revisit the assumption that models necessarily learn similar functions. We conduct a large-scale empirical study of models across hyper-parameters, architectures, frameworks, and datasets. We find that model pairs that diverge more in training methodology display categorically different generalization behavior, producing increasingly uncorrelated errors. We show these models specialize in subdomains of the data, leading to higher ensemble performance: with just 2 models (each with ImageNet accuracy \~76.5\%), we can create ensembles with 83.4\% (+7\% boost). Surprisingly, we find that even significantly low-accuracy models can be used to improve high-accuracy models. Finally, we show diverging training methodology yield representations that capture overlapping (but not supersetting) feature sets which, when combined, lead to increased downstream performance.

Generalized Kernel Thinning

- Raaz Dwivedi, Lester Mackey
- abstract@open-review(Poster): The kernel thinning (KT) algorithm of Dwivedi and Mackey (2021) compresses a probability distribution more effectively than independent sampling by targeting a reproducing kernel Hilbert space (RKHS) and leveraging a less smooth square-root kernel. Here we provide four improvements. First, we show that KT applied directly to the target RKHS yields tighter, dimension-free guarantees for any kernel, any distribution, and any fixed function in the RKHS. Second, we show that, for analytic kernels like Gaussian, inverse multiquadric, and sinc, target KT admits maximum mean discrepancy (MMD) guarantees comparable to or better than those of square-root KT without making explicit use of a square-root kernel. Third, we prove that KT with a fractional power kernel yields better-than-Monte-Carlo MMD guarantees for non-smooth kernels, like Laplace and Matern, that do not have square-roots. Fourth, we establish that KT applied to a sum of the target and power kernels (a procedure we call KT+) simultaneously inherits the improved MMD guarantees of power KT and the tighter individual function guarantees of target KT. In our experiments with target KT and KT+, we witness significant improvements in integration error even in 100 dimensions and when compressing challenging differential equation posteriors.

How Much Can CLIP Benefit Vision-and-Language Tasks?

- Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, Kurt Keutzer
- abstract@open-review(Poster): Most existing Vision-and-Language (V&L) models rely on pre-trained visual encoders, using a relatively small set of manually-annotated data (as compared to web-crawled data), to perceive the visual world. However, it has been observed that large-scale pretraining usually can result in better generalization performance, e.g., CLIP (Contrastive Language-Image Pre-training), trained on a massive amount of image-caption pairs, has shown a strong zero-shot capability on various vision tasks. To further study the advantage brought by CLIP, we propose to use CLIP as the visual encoder in various V&L models in two typical scenarios: 1) plugging CLIP into task-specific fine-tuning; 2) combining CLIP with V&L pre-training and transferring to downstream tasks. We show that CLIP significantly outperforms widely-used visual encoders trained with in-domain annotated data, such as BottomUp-TopDown. We achieve competitive or better results on diverse V&L tasks, while establishing new state-of-the-art results on Visual Question Answering, Visual Entailment, and V&L Navigation tasks.

Large Learning Rate Tames Homogeneity: Convergence and Balancing Effect

- Yuqing Wang, Minshuo Chen, Tuo Zhao, Molei Tao
- abstract@open-review(Poster): Recent empirical advances show that training deep models with large learning rate often improves generalization performance. However, theoretical justifications on the benefits of large learning rate are highly limited, due to challenges in analysis. In this paper, we consider using Gradient Descent (GD) with a large learning rate on a homogeneous matrix factorization problem, i.e., $\min_{X, Y} \|A - XY^\top\|_{\sf F}^2$. We prove a convergence theory for constant large learning rates well beyond $2/L$, where L is the largest eigenvalue of Hessian at the initialization. Moreover, we rigorously establish an implicit bias of GD induced by such a large learning rate, termed `balancing', meaning that magnitudes of X and Y at the limit of GD iterations will be close even if their initialization is significantly unbalanced. Numerical experiments are provided to support our theory.

Demystifying Limited Adversarial Transferability in Automatic Speech Recognition Systems

- Hadi Abdullah, Aditya Karlekar, Vincent Bindschaedler, Patrick Traynor

- abstract@open-review(Poster): The targeted transferability of adversarial samples enables attackers to exploit black-box models in the real-world. The most popular method to produce these adversarial samples is optimization attacks, which have been shown to achieve a high level of transferability in some domains. However, recent research has demonstrated that these attack samples fail to transfer when applied to Automatic Speech Recognition Systems (ASRs). In this paper, we investigate factors preventing this transferability via exhaustive experimentation. To do so, we perform an ablation study on each stage of the ASR pipeline. We discover and quantify six factors (i.e., input type, MFCC, RNN, output type, and vocabulary and sequence sizes) that impact the targeted transferability of optimization attacks against ASRs. Future research can leverage our findings to build ASRs that are more robust to other transferable attack types (e.g., signal processing attacks), or to modify architectures in other domains to reduce their exposure to targeted transferability of optimization attacks.

PipeGCN: Efficient Full-Graph Training of Graph Convolutional Networks with Pipelined Feature Communication

- Cheng Wan, Youjie Li, Cameron R. Wolfe, Anastasios Kyrillidis, Nam Sung Kim, Yingyan Lin
- abstract@open-review(Poster): Graph Convolutional Networks (GCNs) is the state-of-the-art method for learning graph-structured data, and training large-scale GCNs requires distributed training across multiple accelerators such that each accelerator is able to hold a partitioned subgraph. However, distributed GCN training incurs prohibitive overhead of communicating node features and feature gradients among partitions for every GCN layer during each training iteration, limiting the achievable training efficiency and model scalability. To this end, we propose PipeGCN, a simple yet effective scheme that hides the communication overhead by pipelining inter-partition communication with intra-partition computation. It is non-trivial to pipeline for efficient GCN training, as communicated node features/gradients will become stale and thus can harm the convergence, negating the pipeline benefit. Notably, little is known regarding the convergence rate of GCN training with both stale features and stale feature gradients. This work not only provides a theoretical convergence analysis but also finds the convergence rate of PipeGCN to be close to that of the vanilla distributed GCN training without any staleness. Furthermore, we develop a smoothing method to further improve PipeGCN's convergence. Extensive experiments show that PipeGCN can largely boost the training throughput (1.7Ã—~28.5Ã—) while achieving the same accuracy as its vanilla counterpart and existing full-graph training methods. The code is available at https://github.com/RICE-EIC/PipeGCN.

Learning Neural Contextual Bandits through Perturbed Rewards

- Yiling Jia, Weitong ZHANG, Dongruo Zhou, Quanquan Gu, Hongning Wang
- abstract@open-review(Poster): Thanks to the power of representation learning, neural contextual bandit algorithms demonstrate remarkable performance improvement against their classical counterparts. But because their exploration has to be performed in the entire neural network parameter space to obtain nearly optimal regret, the resulting computational cost is prohibitively high.
  We propose to perturb the rewards when updating the neural network to eliminate the need of explicit exploration and the corresponding computational overhead. We prove that a $\tilde{O}(\tilde{d}\sqrt{T})$ regret upper bound is still achievable under standard regularity conditions, where T is the number of rounds of interactions and $\tilde{d}$ is the effective dimension of a neural tangent kernel matrix. Extensive comparisons with several benchmark contextual bandit algorithms, including two recent neural contextual bandit models, demonstrate the effectiveness and computational efficiency of our proposed neural bandit algorithm.

Adversarial Unlearning of Backdoors via Implicit Hypergradient

- Yi Zeng, Si Chen, Won Park, Zhuoqing Mao, Ming Jin, Ruoxi Jia
- abstract@open-review(Poster): We propose a minimax formulation for removing backdoors from a given poisoned model based on a small set of clean data. This formulation encompasses much of prior work on backdoor removal. We propose the Implicit Backdoor Adversarial Unlearning (I-BAU) algorithm to solve the minimax. Unlike previous work, which breaks down the minimax into separate inner and outer problems, our algorithm utilizes the implicit hypergradient to account for the interdependence between inner and outer optimization. We theoretically analyze its convergence and the generalizability of the robustness gained by solving minimax on clean data to unseen test data. In our evaluation, we compare I-BAU with six state-of-art backdoor defenses on eleven backdoor attacks over two datasets and various attack settings, including the common setting where the attacker targets one class as well as important but underexplored settings where multiple classes are targeted. I-BAU's performance is comparable to and most often significantly better than the best baseline. Particularly, its performance is more robust to the variation on triggers, attack settings, poison ratio, and clean data size. Moreover, I-BAU requires less computation to take effect; particularly, it is more than $13\times$ faster than the most efficient baseline in the single-target attack setting. Furthermore, it can remain effective in the extreme case where the defender can only access 100 clean samples---a setting where all the baselines fail to produce acceptable results.

Maximizing Ensemble Diversity in Deep Reinforcement Learning

- Hassam Sheikh, Mariano Phielipp, Ladislau Boloni
- abstract@open-review(Poster): Modern deep reinforcement learning (DRL) has been successful in solving a range of challenging sequential decision-making problems. Most of these algorithms use an ensemble of neural networks as their backbone structure and benefit from the diversity among the neural networks to achieve optimal results. Unfortunately, the members of the ensemble can converge to the same point either the parametric space or representation space during the training phase, therefore, losing all the leverage of an ensemble. In this paper, we describe Maximize Ensemble Diversity in Reinforcement Learning (MED-RL), a set of regularization methods inspired from the economics and consensus optimization to improve diversity in the ensemble-based deep reinforcement learning methods by encouraging inequality between the networks during training. We integrated MED-RL in five of the most common ensemble-based deep RL algorithms for both continuous and discrete control tasks and evaluated on six Mujoco environments and six Atari games. Our results show that MED-RL augmented algorithms outperform their un-regularized counterparts significantly and in some cases achieved more than 300$\%$ in performance gains.

Graph Neural Networks with Learnable Structural and Positional Representations

- Vijay Prakash Dwivedi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, Xavier Bresson
- abstract@open-review(Poster): Graph neural networks (GNNs) have become the standard learning architectures for graphs. GNNs have been applied to numerous domains ranging from quantum chemistry, recommender systems to knowledge graphs and natural language processing. A major issue with arbitrary graphs is the absence of canonical positional information of nodes, which decreases the representation power of GNNs to distinguish e.g. isomorphic nodes and other graph symmetries. An approach to tackle this issue is to introduce Positional Encoding (PE) of nodes, and inject it into the input layer, like in Transformers. Possible graph PE are Laplacian eigenvectors. In this work, we propose to decouple structural and positional representations to make easy for the network to learn these two essential properties. We introduce a novel generic architecture which we call \texttt{LSPE} (Learnable Structural and Positional Encodings). We investigate several sparse and fully-connected (Transformer-like) GNNs, and observe a performance increase for molecular datasets, from 1.79% up to 64.14% when considering learnable PE for both GNN classes.

Zero-Shot Self-Supervised Learning for MRI Reconstruction

- Burhaneddin Yaman, Seyed Amir Hossein Hosseini, Mehmet Akcakaya
- abstract@open-review(Poster): Deep learning (DL) has emerged as a powerful tool for accelerated MRI reconstruction, but often necessitates a database of fully-sampled measurements for training. Recent self-supervised and unsupervised learning approaches enable training without fully-sampled data. However, a database of undersampled measurements may not be available in many scenarios, especially for scans involving contrast or translational

acquisitions in development. Moreover, recent studies show that database-trained models may not generalize well when the unseen measurements differ in terms of sampling pattern, acceleration rate, SNR, image contrast, and anatomy. Such challenges necessitate a new methodology to enable subject-specific DL MRI reconstruction without external training datasets, since it is clinically imperative to provide high-quality reconstructions that can be used to identify lesions/disease for $\textit{every individual}$. In this work, we propose a zero-shot self-supervised learning approach to perform subject-specific accelerated DL MRI reconstruction to tackle these issues. The proposed approach partitions the available measurements from a single scan into three disjoint sets. Two of these sets are used to enforce data consistency and define loss during training for self-supervision, while the last set serves to self-validate, establishing an early stopping criterion. In the presence of models pre-trained on a database with different image characteristics, we show that the proposed approach can be combined with transfer learning for faster convergence time and reduced computational complexity.

Policy Smoothing for Provably Robust Reinforcement Learning

- Aounon Kumar, Alexander Levine, Soheil Feizi
- abstract@open-review(Poster): The study of provable adversarial robustness for deep neural networks (DNNs) has mainly focused on $\textit{static}$ supervised learning tasks such as image classification. However, DNNs have been used extensively in real-world $\textit{adaptive}$ tasks such as reinforcement learning (RL), making such systems vulnerable to adversarial attacks as well. Prior works in provable robustness in RL seek to certify the behaviour of the victim policy at every time-step against a non-adaptive adversary using methods developed for the static setting. But in the real world, an RL adversary can infer the defense strategy used by the victim agent by observing the states, actions, etc. from previous time-steps and adapt itself to produce stronger attacks in future steps (e.g., by focusing more on states critical to the agent's performance). We present an efficient procedure, designed specifically to defend against an adaptive RL adversary, that can directly certify the total reward without requiring the policy to be robust at each time-step. Focusing on randomized smoothing based defenses, our main theoretical contribution is to prove an $\textit{adaptive version}$ of the Neyman-Pearson Lemma -- a key lemma for smoothing-based certificates -- where the adversarial perturbation at a particular time can be a stochastic function of current and previous observations and states as well as previous actions. Building on this result, we propose $\textit{policy smoothing}$ where the agent adds a Gaussian noise to its observation at each time-step before passing it through the policy function. Our robustness certificates guarantee that the final total reward obtained by policy smoothing remains above a certain threshold, even though the actions at intermediate time-steps may change under the attack. We show that our certificates are $\textit{tight}$ by constructing a worst-case scenario that achieves the bounds derived in our analysis. Our experiments on various environments like Cartpole, Pong, Freeway and Mountain Car show that our method can yield meaningful robustness guarantees in practice.

The Close Relationship Between Contrastive Learning and Meta-Learning

- Renkun Ni, Manli Shu, Hossein Souri, Micah Goldblum, Tom Goldstein
- abstract@open-review(Poster): Contrastive learning has recently taken off as a paradigm for learning from unlabeled data. In this paper, we discuss the close relationship between contrastive learning and meta-learning under a certain task distribution. We complement this observation by showing that established meta-learning methods, such as Prototypical Networks, achieve comparable performance to SimCLR when paired with this task distribution. This relationship can be leveraged by taking established techniques from meta-learning, such as task-based data augmentation, and showing that they benefit contrastive learning as well. These tricks also benefit state-of-the-art self-supervised learners without using negative pairs such as BYOL, which achieves 94.6\% accuracy on CIFAR-10 using a self-supervised ResNet-18 feature extractor trained with our meta-learning tricks. We conclude that existing advances designed for contrastive learning or meta-learning can be exploited to benefit the other, and it is better for contrastive learning researchers to take lessons from the meta-learning literature (and vice-versa) than to reinvent the wheel.

Towards Understanding Generalization via Decomposing Excess Risk Dynamics

- Jiaye Teng, Jianhao Ma, Yang Yuan
- abstract@open-review(Poster): Generalization is one of the fundamental issues in machine learning. However, traditional techniques like uniform convergence may be unable to explain generalization under overparameterization \citep{nagarajan2019uniform}. As alternative approaches, techniques based on stability analyze the training dynamics and derive algorithm-dependent generalization bounds. Unfortunately, the stability-based bounds are still far from explaining the surprising generalization in deep learning since neural networks usually suffer from unsatisfactory stability. This paper proposes a novel decomposition framework to improve the stability-based bounds via a more fine-grained analysis of the signal and noise, inspired by the observation that neural networks converge relatively slowly when fitting noise (which indicates better stability). Concretely, we decompose the excess risk dynamics and apply the stability-based bound only on the noise component. The decomposition framework performs well in both linear regimes (overparameterized linear regression) and non-linear regimes (diagonal matrix recovery). Experiments on neural networks verify the utility of the decomposition framework.

Graph Auto-Encoder via Neighborhood Wasserstein Reconstruction

- Mingyue Tang, Pan Li, Carl Yang
- abstract@open-review(Poster): Graph neural networks (GNNs) have drawn significant research attention recently, mostly under the setting of semi-supervised learning. When task-agnostic representations are preferred or supervision is simply unavailable, the auto-encoder framework comes in handy with a natural graph reconstruction objective for unsupervised GNN training. However, existing graph auto-encoders are designed to reconstruct the direct links, so GNNs trained in this way are only optimized towards proximity-oriented graph mining tasks, and will fall short when the topological structures matter. In this work, we revisit the graph encoding process of GNNs which essentially learns to encode the neighborhood information of each node into an embedding vector, and propose a novel graph decoder to reconstruct the entire neighborhood information regarding both proximity and structure via Neighborhood Wasserstein Reconstruction (NWR). Specifically, from the GNN embedding of each node, NWR jointly predicts its node degree and neighbor feature distribution, where the distribution prediction adopts an optimal-transport loss based on the Wasserstein distance. Extensive experiments on both synthetic and real-world network datasets show that the unsupervised node representations learned with NWR have much more advantageous in structure-oriented graph mining tasks, while also achieving competitive performance in proximity-oriented ones.

FairCal: Fairness Calibration for Face Verification

- Tiago Salvador, Stephanie Cairns, Vikram Voleti, Noah Marshall, Adam M Oberman
- abstract@open-review(Poster): Despite being widely used, face recognition models suffer from bias: the probability of a false positive (incorrect face match) strongly depends on sensitive attributes such as the ethnicity of the face. As a result, these models can disproportionately and negatively impact minority groups, particularly when used by law enforcement. The majority of bias reduction methods have several drawbacks: they use an end-to-end retraining approach, may not be feasible due to privacy issues, and often reduce accuracy. An alternative approach is post-processing methods that build fairer decision classifiers using the features of pre-trained models, thus avoiding the cost of retraining. However, they still have drawbacks: they reduce accuracy (AGENDA, FTC), or require retuning for different false positive rates (FSN). In this work, we introduce the Fairness Calibration (FairCal) method, a post-training approach that simultaneously: (i) increases model accuracy (improving the state-of-the-art), (ii) produces fairly-calibrated probabilities, (iii) significantly reduces the gap in the false positive rates, (iv) does not require knowledge of the sensitive attribute, and (v) does not require retraining, training an additional model or retuning. We apply it to the task of Face Verification, and obtain state-of-the-art results with all the above advantages.

Cross-Lingual Transfer with Class-Weighted Language-Invariant Representations

- Ruicheng Xian, Heng Ji, Han Zhao
- abstract@open-review(Poster): Recent advances in neural modeling have produced deep multilingual language models capable of extracting cross-lingual knowledge from non-parallel texts and enabling zero-shot downstream transfer. While their success is often attributed to shared representations, quantitative analyses are limited. Towards a better understanding, through empirical analyses, we show that the invariance of feature representations across languagesâ€”an effect of shared representationsâ€”strongly correlates with transfer performance. We also observe that distributional shifts in class priors between source and target language task data negatively affect performance, a largely overlooked issue that could cause negative transfer with existing unsupervised approaches. Based on these findings, we propose and evaluate a method for unsupervised transfer, called importance-weighted domain alignment (IWDA), that performs representation alignment with prior shift estimation and correction using unlabeled target language task data. Experiments demonstrate its superiority under large prior shifts, and show further performance gains when combined with existing semi-supervised learning techniques.

ComPhy: Compositional Physical Reasoning of Objects and Events from Videos

- Zhenfang Chen, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B. Tenenbaum, Chuang Gan
- abstract@open-review(Poster): Objects' motions in nature are governed by complex interactions and their properties. While some properties, such as shape and material, can be identified via the object's visual appearances, others like mass and electric charge are not directly visible. The compositionality between the visible and hidden properties poses unique challenges for AI models to reason from the physical world, whereas humans can effortlessly infer them with limited observations. Existing studies on video reasoning mainly focus on visually observable elements such as object appearance, movement, and contact interaction. In this paper, we take an initial step to highlight the importance of inferring the hidden physical properties not directly observable from visual appearances, by introducing the Compositional Physical Reasoning (ComPhy) dataset. For a given set of objects, ComPhy includes few videos of them moving and interacting under different initial conditions. The model is evaluated based on its capability to unravel the compositional hidden properties, such as mass and charge, and use this knowledge to answer a set of questions posted on one of the videos. Evaluation results of several state-of-the-art video reasoning models on ComPhy show unsatisfactory performance as they fail to capture these hidden properties. We further propose an oracle neural-symbolic framework named Compositional Physics Learner (CPL), combining visual perception, physical property learning, dynamic prediction, and symbolic execution into a unified framework. CPL can effectively identify objects' physical properties from their interactions and predict their dynamics to answer questions.

An Information Fusion Approach to Learning with Instance-Dependent Label Noise

- Zhimeng Jiang, Kaixiong Zhou, Zirui Liu, Li Li, Rui Chen, Soo-Hyun Choi, Xia Hu
- abstract@open-review(Poster): Instance-dependent label noise (IDN) widely exists in real-world datasets and usually misleads the training of deep neural networks. Noise transition matrix (NTM) (i.e., the probability that clean labels flip into noisy labels) is used to characterize the label noise and can be adopted to bridge the gap between clean and noisy underlying data distributions. However, most instances are long-tail, i.e., the number of occurrences of each instance is usually limited, which leads to the gap between the underlying distribution and the empirical distribution. Therefore, the genuine problem caused by IDN is \emph{empirical}, instead of underlying, \emph{data distribution mismatch} during training. To directly tackle the empirical distribution mismatch problem, we propose \emph{posterior transition matrix} (PTM) to posteriorly model label noise given limited observed noisy labels, which achieves \emph{statistically consistent classifiers}. Note that even if an instance is corrupted by the same NTM, the intrinsic randomness incurs different noisy labels, and thus requires different correction methods. Motivated by this observation, we propose an \textbf{I}nformation \textbf{F}usion (IF) approach to fine-tune the NTM based on the estimated PTM. Specifically, we adopt the noisy labels and model predicted probabilities to estimate the PTM and then correct the NTM in \emph{forward propagation}. Empirical evaluations on synthetic and real-world datasets demonstrate that our method is superior to the state-of-the-art approaches, and achieves more stable training for instance-dependent label noise.

On Redundancy and Diversity in Cell-based Neural Architecture Search

- Xingchen Wan, Binxin Ru, Pedro M EsperanÃ§a, Zhenguo Li
- abstract@open-review(Poster): Searching for the architecture cells is a dominant paradigm in NAS. However, little attention has been devoted to the analysis of the cell-based search spaces even though it is highly important for the continual development of NAS. In this work, we conduct an empirical post-hoc analysis of architectures from the popular cell-based search spaces and find that the existing search spaces contain a high degree of redundancy: the architecture performance is less sensitive to changes at large parts of the cells, and universally adopted design rules, like the explicit search for a reduction cell, significantly increase the complexities but have very limited impact on the performance. Across architectures found by a diverse set of search strategies, we consistently find that the parts of the cells that do matter for architecture performance often follow similar and simple patterns. By constraining cells to include these patterns, randomly sampled architectures can match or even outperform the state of the art. These findings cast doubts into our ability to discover truly novel architectures in the existing cell-based search spaces and, inspire our suggestions for improvement to guide future NAS research. Code is available at https://github.com/xingchenwan/cell-based-NAS-analysis.

Deep Learning without Shortcuts: Shaping the Kernel with Tailored Rectifiers

- Guodong Zhang, Aleksandar Botev, James Martens
- abstract@open-review(Poster): Training very deep neural networks is still an extremely challenging task. The common solution is to use shortcut connections and normalization layers, which are both crucial ingredients in the popular ResNet architecture. However, there is strong evidence to suggest that ResNets behave more like ensembles of shallower networks than truly deep ones. Recently, it was shown that deep vanilla networks (i.e.~networks without normalization layers or shortcut connections) can be trained as fast as ResNets by applying certain transformations to their activation functions. However, this method (called Deep Kernel Shaping) isn't fully compatible with ReLUs, and produces networks that overfit significantly more than ResNets on ImageNet. In this work, we rectify this situation by developing a new type of transformation that is fully compatible with a variant of ReLUs -- Leaky ReLUs. We show in experiments that our method, which introduces negligible extra computational cost, achieves validation accuracies with deep vanilla networks that are competitive with ResNets (of the same width/depth), and significantly higher than those obtained with the Edge of Chaos (EOC) method. And unlike with EOC, the validation accuracies we obtain do not get worse with depth.

Variational autoencoders in the presence of low-dimensional data: landscape and implicit bias

- Frederic Koehler, Viraj Mehta, Chenghui Zhou, Andrej Risteski
- abstract@open-review(Poster): Variational Autoencoders (VAEs) are one of the most commonly used generative models, particularly for image data. A prominent difficulty in training VAEs is data that is supported on a lower dimensional manifold. Recent work by Dai and Wipf (2020) proposes a two-stage training algorithm for VAEs, based on a conjecture that in standard VAE training the generator will converge to a solution with 0 variance which is correctly supported on the ground truth manifold. They gave partial support for this conjecture by showing that some optima of the VAE loss do satisfy this property, but did not analyze the training dynamics. In this paper, we show that for linear encoders/decoders, the conjecture is trueâ€”that is the VAE training does recover a generator with support equal to the ground truth manifoldâ€”and does so due to an implicit bias of gradient descent rather than merely the VAE loss itself. In the nonlinear case, we show that VAE training frequently learns a higher-dimensional manifold which is a superset of the ground truth manifold.

No Parameters Left Behind: Sensitivity Guided Adaptive Learning Rate for Training Large Transformer Models

- Chen Liang, Haoming Jiang, Simiao Zuo, Pengcheng He, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, Tuo Zhao
- abstract@open-review(Poster): Recent research has shown the existence of significant redundancy in large Transformer models. One can prune the redundant parameters without significantly sacrificing the generalization performance. However, we question whether the redundant parameters could have contributed more if they were properly trained. To answer this question, we propose a novel training strategy that encourages all parameters to be trained sufficiently. Specifically, we adaptively adjust the learning rate for each parameter according to its sensitivity, a robust gradient-based measure reflecting this parameter's contribution to the model performance. A parameter with low sensitivity is redundant, and we improve its fitting by increasing its learning rate. In contrast, a parameter with high sensitivity is well-trained, and we regularize it by decreasing its learning rate to prevent further overfitting. We conduct extensive experiments on natural language understanding, neural machine translation, and image classification to demonstrate the effectiveness of the proposed schedule. Analysis shows that the proposed schedule indeed reduces the redundancy and improves generalization performance.

SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations

- Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, Stefano Ermon
- abstract@open-review(Poster): Guided image synthesis enables everyday users to create and edit photo-realistic images with minimum effort. The key challenge is balancing faithfulness to the user inputs (e.g., hand-drawn colored strokes) and realism of the synthesized images. Existing GAN-based methods attempt to achieve such balance using either conditional GANs or GAN inversions, which are challenging and often require additional training data or loss functions for individual applications. To address these issues, we introduce a new image synthesis and editing method, Stochastic Differential Editing (SDEdit), based on a diffusion model generative prior, which synthesizes realistic images by iteratively denoising through a stochastic differential equation (SDE). Given an input image with user guide in a form of manipulating RGB pixels, SDEdit first adds noise to the input, then subsequently denoises the resulting image through the SDE prior to increase its realism. SDEdit does not require task-specific training or inversions and can naturally achieve the balance between realism and faithfulness. SDEdit outperforms state-of-the-art GAN-based methods by up to 98.09% on realism and 91.72% on overall satisfaction scores, according to a human perception study, on multiple tasks, including stroke-based image synthesis and editing as well as image compositing.

Post hoc Explanations may be Ineffective for Detecting Unknown Spurious Correlation

- Julius Adebayo, Michael Muelly, Harold Abelson, Been Kim
- abstract@open-review(Poster): We investigate whether three types of post hoc model explanationsâ€“feature attribution, concept activation, and training point rankingâ€“are effective for detecting a modelâ€™s reliance on spurious signals in the training data. Specifically, we consider the scenario where the spurious signal to be detected is unknown, at test-time, to the user of the explanation method. We design an empirical methodology that uses semi-synthetic datasets along with pre-specified spurious artifacts to obtain models that verifiably rely on these spurious training signals. We then provide a suite of metrics that assess an explanation methodâ€™s reliability for spurious signal detection under various conditions. We find that the post hoc explanation methods tested are ineffective when the spurious artifact is unknown at test-time especially for non-visible artifacts like a background blur. Further, we find that feature attribution methods are susceptible to erroneously indicating dependence on spurious signals even when the model being explained does not rely on spurious artifacts. This finding casts doubt on the utility of these approaches, in the hands of a practitioner, for detecting a modelâ€™s reliance on spurious signals.

Generalizing Few-Shot NAS with Gradient Matching

- Shoukang Hu, Ruochen Wang, Lanqing HONG, Zhenguo Li, Cho-Jui Hsieh, Jiashi Feng
- abstract@open-review(Poster): Efficient performance estimation of architectures drawn from large search spaces is essential to Neural Architecture Search. One-Shot methods tackle this challenge by training one supernet to approximate the performance of every architecture in the search space via weight-sharing, thereby drastically reducing the search cost. However, due to coupled optimization between child architectures caused by weight-sharing, One-Shot supernet's performance estimation could be inaccurate, leading to degraded search outcomes. To address this issue, Few-Shot NAS reduces the level of weight-sharing by splitting the One-Shot supernet into multiple separated sub-supernets via edge-wise (layer-wise) exhaustive partitioning. Since each partition of the supernet is not equally important, it necessitates the design of a more effective splitting criterion. In this work, we propose a gradient matching score (GM) that leverages gradient information at the shared weight for making informed splitting decisions. Intuitively, gradients from different child models can be used to identify whether they agree on how to update the shared modules, and subsequently to decide if they should share weight. Compared with exhaustive partitioning, the proposed criterion significantly reduces the branching factor per edge. This allows us to split more edges (layers) for a given budget, resulting in substantially improved performance as NAS search spaces usually include dozens of edges (layers). Extensive empirical evaluations of the proposed method on a wide range of search spaces (NASBench-201, DARTS, MobileNet Space), datasets (cifar10, cifar100, ImageNet) and search algorithms (DARTS, SNAS, RSPS, ProxylessNAS, OFA) demonstrate that it significantly outperforms its Few-Shot counterparts while surpassing previous comparable methods in terms of the accuracy of derived architectures. Our code is available at https://github.com/skhu101/GM-NAS.

The Unreasonable Effectiveness of Random Pruning: Return of the Most Naive Baseline for Sparse Training

- Shiwei Liu, Tianlong Chen, Xiaohan Chen, Li Shen, Decebal Constantin Mocanu, Zhangyang Wang, Mykola Pechenizkiy
- abstract@open-review(Poster): Random pruning is arguably the most naive way to attain sparsity in neural networks, but has been deemed uncompetitive by either post-training pruning or sparse training. In this paper, we focus on sparse training and highlight a perhaps counter-intuitive finding, that random pruning at initialization can be quite powerful for the sparse training of modern neural networks. Without any delicate pruning criteria or carefully pursued sparsity structures, we empirically demonstrate that sparsely training a randomly pruned network from scratch can match the performance of its dense equivalent. There are two key factors that contribute to this revival: (i) $the network sizes matter$: as the original dense networks grow wider and deeper, the performance of training a randomly pruned sparse network will quickly grow to matching that of its dense equivalent, even at high sparsity ratios; (ii) $appropriate layer-wise sparsity ratios$ can be pre-chosen for sparse training, which shows to be another important performance booster. Simple as it looks, a randomly pruned subnetwork of Wide ResNet-50 can be sparsely trained to outperforming a dense Wide ResNet-50, on ImageNet. We also observed such randomly pruned networks outperform dense counterparts in other favorable aspects, such as out-of-distribution detection, uncertainty estimation, and adversarial robustness. Overall, our results strongly suggest there is larger-than-expected room for sparse training at scale, and the benefits of sparsity might be more universal beyond carefully designed pruning. Our source code can be found at https://github.com/VITA-Group/Random_Pruning.

switch-GLAT: Multilingual Parallel Machine Translation Via Code-Switch Decoder

- Zhenqiao Song, Hao Zhou, Lihua Qian, Jingjing Xu, Shanbo Cheng, Mingxuan Wang, Lei Li
- abstract@open-review(Poster): Multilingual machine translation aims to develop a single model for multiple language directions. However, existing multilingual models based on Transformer are limited in terms of both translation performance and inference speed. In this paper, we propose switch-GLAT, a non-autoregressive multilingual machine translation model with a code-switch decoder. It can generate contextual code-switched translations for a given source sentence, and perform code-switch back-translation, greatly boosting multilingual translation performance. In addition, its inference is highly efficient thanks to its parallel decoder. Experiments show that our proposed switch-GLAT outperform the multilingual Transformer with as much as 0.74 BLEU improvement and 6.2x faster decoding speed in inference.

DictFormer: Tiny Transformer with Shared Dictionary

- Qian Lou, Ting Hua, Yen-Chang Hsu, Yilin Shen, Hongxia Jin
- abstract@open-review(Poster): We introduce DictFormer with the efficient shared dictionary to provide a compact, fast, and accurate transformer model. DictFormer significantly reduces the redundancy in the transformer's parameters by replacing the prior transformer's parameters with a compact, shared dictionary, few unshared coefficients, and indices. Also, DictFormer enables faster computations since expensive weights multiplications are converted into cheap shared look-ups on dictionary and few linear projections. Training dictionary and coefficients are not trivial since indices used for looking up dictionary are not differentiable. We adopt a sparse-constraint training with $l_1\,\,norm$ relaxation to learn coefficients and indices in DictFormer. DictFormer is flexible to support different model sizes by dynamically changing dictionary size. Compared to existing lightweight Transformers, DictFormer consistently reduces model size over Transformer on multiple tasks, e.g., machine translation, abstractive summarization, and language modeling. Extensive experiments show that DictFormer reduces $3.6\times$ to $8.9\times$ model size with similar accuracy over multiple tasks, compared to Transformer.

Training Transition Policies via Distribution Matching for Complex Tasks

- JU-SEUNG BYUN, Andrew Perrault
- abstract@open-review(Poster): Humans decompose novel complex tasks into simpler ones to exploit previously learned skills. Analogously, hierarchical reinforcement learning seeks to leverage lower-level policies for simple tasks to solve complex ones. However, because each lower-level policy induces a different distribution of states, transitioning from one lower-level policy to another may fail due to an unexpected starting state. We introduce transition policies that smoothly connect lower-level policies by producing a distribution of states and actions that matches what is expected by the next policy. Training transition policies is challenging because the natural reward signal---whether the next policy can execute its subtask successfully---is sparse. By training transition policies via adversarial inverse reinforcement learning to match the distribution of expected states and actions, we avoid relying on task-based reward. To further improve performance, we use deep Q-learning with a binary action space to determine when to switch from a transition policy to the next pre-trained policy, using the success or failure of the next subtask as the reward. Although the reward is still sparse, the problem is less severe due to the simple binary action space. We demonstrate our method on continuous bipedal locomotion and arm manipulation tasks that require diverse skills. We show that it smoothly connects the lower-level policies, achieving higher success rates than previous methods that search for successful trajectories based on a reward function, but do not match the state distribution.

GDA-AM: ON THE EFFECTIVENESS OF SOLVING MIN-IMAX OPTIMIZATION VIA ANDERSON MIXING

- Huan He, Shifan Zhao, Yuanzhe Xi, Joyce Ho, Yousef Saad
- abstract@open-review(Poster): Many modern machine learning algorithms such as generative adversarial networks (GANs) and adversarial training can be formulated as minimax optimization.Gradient descent ascent (GDA) is the most commonly used algorithm due to its simplicity. However, GDA can converge to non-optimal minimax points. We propose a new minimax optimization framework,GDA-AM, that views the GDA dynamics as a fixed-point iteration and solves it using Anderson Mixing to converge to the local minimax. It addresses the diverging issue of simultaneous GDA and accelerates the convergence of alternating GDA. We show theoretically that the algorithm can achieve global convergence for bilinear problems under mildconditions. We also empirically show that GDA-AM solves a variety of minimax problems and improves GAN training on several datasets

On feature learning in neural networks with global convergence guarantees

- Zhengdao Chen, Eric Vanden-Eijnden, Joan Bruna
- abstract@open-review(Poster): We study the gradient flow optimization of over-parameterized neural networks (NNs) in a setup that allows feature learning while admitting non-asymptotic global convergence guarantees. First, we prove that for wide shallow NNs under the mean-field (MF) scaling and with a general class of activation functions, when the input dimension is at least the size of the training set, the training loss converges to zero at a linear rate under gradient flow. Building upon this analysis, we study a model of wide multi-layer NNs with random and untrained weights in earlier layers, and also prove a linear-rate convergence of the training loss to zero, regardless of the input dimension. We also show empirically that, unlike in the Neural Tangent Kernel (NTK) regime, our multi-layer model exhibits feature learning and can achieve better generalization performance than its NTK counterpart.

The Three Stages of Learning Dynamics in High-dimensional Kernel Methods

- Nikhil Ghosh, Song Mei, Bin Yu
- abstract@open-review(Poster): To understand how deep learning works, it is crucial to understand the training dynamics of neural networks. Several interesting hypotheses about these dynamics have been made based on empirically observed phenomena, but there exists a limited theoretical understanding of when and why such phenomena occur.

In this paper, we consider the training dynamics of gradient flow on kernel least-squares objectives, which is a limiting dynamics of SGD trained neural networks. Using precise high-dimensional asymptotics, we characterize the dynamics of the fitted model in two â€œworldsâ€ : in the Oracle World the model is trained on the population distribution and in the Empirical World the model is trained on an i.i.d finite dataset. We show that under mild conditions on the kernel and L^2 target regression function the training dynamics have three stages that are based on the behaviors of the models in the two worlds. Our theoretical results also mathematically formalize some interesting deep learning phenomena. Specifically, in our setting we show that SGD progressively learns more complex functions and that there is a "deep bootstrap" phenomenon: during the second stage, the test error of both worlds remain close despite the empirical training error being much smaller. Finally, we give a concrete example comparing the dynamics of two different kernels which shows that faster training is not necessary for better generalization.

When Can We Learn General-Sum Markov Games with a Large Number of Players Sample-Efficiently?

- Ziang Song, Song Mei, Yu Bai
- abstract@open-review(Poster): Multi-agent reinforcement learning has made substantial empirical progresses in solving games with a large number of players. However, theoretically, the best known sample complexity for finding a Nash equilibrium in general-sum games scales exponentially in the number of players due to the size of the joint action space, and there is a matching exponential lower bound. This paper investigates what learning goals admit better sample complexities in the setting of m-player general-sum Markov games with H steps, S states, and A_i actions per player. First, we design algorithms for learning an ϵ-Coarse Correlated Equilibrium (CCE) in $\widetilde{\mathcal{O}}(H^5S\max_{i\le m} A_i / \epsilon^2)$ episodes, and an ϵ-Correlated Equilibrium (CE) in $\widetilde{\mathcal{O}}(H^6S\max_{i\le m} A_i^2 / \epsilon^2)$ episodes. This is the first line of results for learning CCE and CE with sample complexities polynomial in $\max_{i\le m} A_i$. Our algorithm for learning CE integrates an adversarial bandit subroutine which minimizes a weighted swap regret, along with several novel designs in the outer loop. Second, we consider the important special case of Markov Potential Games, and design an algorithm that learns an ϵ-approximate Nash equilibrium within $\widetilde{\mathcal{O}}(S\sum_{i\le m} A_i / \epsilon^3)$ episodes (when only highlighting the dependence on S, A_i, and ϵ), which only depends linearly in $\sum_{i\le m} A_i$ and significantly improves over the existing efficient algorithm in the ϵ dependence. Overall, our results shed light on what equilibria or structural assumptions on the game may enable sample-efficient learning with many players.

Neural Networks as Kernel Learners: The Silent Alignment Effect

- Alexander Atanasov, Blake Bordelon, Cengiz Pehlevan
- abstract@open-review(Poster): Neural networks in the lazy training regime converge to kernel machines. Can neural networks in the rich feature learning regime learn a kernel machine with a data-dependent kernel? We demonstrate that this can indeed happen due to a phenomenon we term silent alignment, which requires that the tangent kernel of a network evolves in eigenstructure while small and before the loss appreciably decreases, and grows only in overall scale afterwards. We show that such an effect takes place in homogenous neural networks with small initialization and whitened data. We provide an analytical treatment of this effect in the linear network case. In general, we find that the kernel develops a low-rank contribution in the early phase of training, and then evolves in overall scale, yielding a function equivalent to a kernel regression solution with the final network's tangent kernel. The early spectral learning of the kernel depends on the depth. We also demonstrate that non-whitened data can weaken the silent alignment effect.

Learning Object-Oriented Dynamics for Planning from Text

- Guiliang Liu, Ashutosh Adhikari, Amir-massoud Farahmand, Pascal Poupart
- abstract@open-review(Poster): The advancement of dynamics models enables model-based planning in complex environments. Existing dynamics models commonly study image-based games with fully observable states. Generalizing these models to Text-Based Games (TBGs), which commonly describe the partially observable states with noisy text observations, is challenging. In this work, we propose an Object-Oriented Text Dynamics (OOTD) model that enables planning algorithms to solve decision-making problems in text domains. OOTD predicts a memory graph that dynamically remembers the history of object observations and filters object-irrelevant information. To facilitate the robustness of dynamics, our OOTD model identifies the objects influenced by input actions and predicts the belief of object states with independently parameterized transition layers. We develop variational objectives under the object-supervised and self-supervised settings to model the stochasticity of predicted dynamics. Empirical results show OOTD-based planner significantly outperforms model-free baselines in terms of sample efficiency and running scores.

An Operator Theoretic View On Pruning Deep Neural Networks

- William T Redman, MARIA FONOBEROVA, Ryan Mohr, Yannis Kevrekidis, Igor Mezic
- abstract@open-review(Poster): The discovery of sparse subnetworks that are able to perform as well as full models has found broad applied and theoretical interest. While many pruning methods have been developed to this end, the naÃ¯ve approach of removing parameters based on their magnitude has been found to be as robust as more complex, state-of-the-art algorithms. The lack of theory behind magnitude pruning's success, especially pre-convergence, and its relation to other pruning methods, such as gradient based pruning, are outstanding open questions in the field that are in need of being addressed. We make use of recent advances in dynamical systems theory, namely Koopman operator theory, to define a new class of theoretically motivated pruning algorithms. We show that these algorithms can be equivalent to magnitude and gradient based pruning, unifying these seemingly disparate methods, and find that they can be used to shed light on magnitude pruning's performance during the early part of training.

Capacity of Group-invariant Linear Readouts from Equivariant Representations: How Many Objects can be Linearly Classified Under All Possible Views?

- Matthew Farrell, Blake Bordelon, Shubhendu Trivedi, Cengiz Pehlevan
- abstract@open-review(Poster): Equivariance has emerged as a desirable property of representations of objects subject to identity-preserving transformations that constitute a group, such as translations and rotations. However, the expressivity of a representation constrained by group equivariance is still not fully understood. We address this gap by providing a generalization of Cover's Function Counting Theorem that quantifies the number of linearly separable and group-invariant binary dichotomies that can be assigned to equivariant representations of objects. We find that the fraction of separable dichotomies is determined by the dimension of the space that is fixed by the group action. We show how this relation extends to operations such as convolutions, element-wise nonlinearities, and global and local pooling. While other operations do not change the fraction of separable dichotomies, local pooling decreases the fraction, despite being a highly nonlinear operation. Finally, we test our theory on intermediate representations of randomly initialized and fully trained convolutional neural networks and find perfect agreement.

Tuformer: Data-driven Design of Transformers for Improved Generalization or Efficiency

- Xiaoyu Liu, Jiahao Su, Furong Huang
- abstract@open-review(Poster): Transformers are neural network architectures that achieve remarkable performance in many areas. However, the core component of Transformers, multi-head self-attention (MHSA), is mainly derived from heuristics, and the interactions across its components are not well understood. To address the problem, we first introduce a mathematically rigorous and yet intuitive tensor diagram representation of MHSA. Guided by tensor diagram representations, we propose a novel design, namely Tunable Transformers (Tuformers), by allowing data-driven weights across heads, whereas MHSA adopts pre-defined and fixed weights across heads, as will be explained in our paper. Tuformers naturally reveal a flexible design space that a user, depending on the needs, can choose a structure that has either improved performance (generalization error) or higher model efficiency. Any pre-trained Transformer can be an initialization of the corresponding Tuformer with trainable number of heads for efficient training and fine-tuning. Tuformers universally outperform Transformers on various tasks across multiple domains under a wide range of model sizes.

Learning Weakly-supervised Contrastive Representations

- Yao-Hung Hubert Tsai, Tianqin Li, Weixin Liu, Peiyuan Liao, Ruslan Salakhutdinov, Louis-Philippe Morency
- abstract@open-review(Poster): We argue that a form of the valuable information provided by the auxiliary information is its implied data clustering information. For instance, considering hashtags as auxiliary information, we can hypothesize that an Instagram image will be semantically more similar with the same hashtags. With this intuition, we present a two-stage weakly-supervised contrastive learning approach. The first stage is to cluster data according to its auxiliary information. The second stage is to learn similar representations within the same cluster and dissimilar representations for data from different clusters. Our empirical experiments suggest the following three contributions. First, compared to conventional self-supervised representations, the auxiliary-information-infused representations bring the performance closer to the supervised representations, which use direct downstream labels as supervision signals. Second, our approach performs the best in most cases, when comparing our approach with other baseline representation learning methods that also leverage auxiliary data information. Third, we show that our approach also works well with unsupervised constructed clusters (e.g., no auxiliary information), resulting in a strong unsupervised representation learning approach.

Encoding Weights of Irregular Sparsity for Fixed-to-Fixed Model Compression

- Bae Seong Park, Se Jung Kwon, Daehwan Oh, Byeongwook Kim, Dongsoo Lee
- abstract@open-review(Poster): Even though fine-grained pruning techniques achieve a high compression ratio, conventional sparsity representations (such as CSR) associated with irregular sparsity degrade parallelism significantly. Practical pruning methods, thus, usually lower pruning rates (by structured pruning) to improve parallelism. In this paper, we study fixed-to-fixed (lossless) encoding architecture/algorithm to support fine-grained pruning methods such that sparse neural networks can be stored in a highly regular structure. We first estimate the maximum compression ratio of encoding-based compression using entropy. Then, as an effort to push the compression ratio to the theoretical maximum (by entropy), we propose a sequential fixed-to-

fixed encoding scheme. We demonstrate that our proposed compression scheme achieves almost the maximum compression ratio for the Transformer and ResNet-50 pruned by various fine-grained pruning methods.

An Experimental Design Perspective on Model-Based Reinforcement Learning

- Viraj Mehta, Biswajit Paria, Jeff Schneider, Stefano Ermon, Willie Neiswanger
- abstract@open-review(Poster): In many practical applications of RL, it is expensive to observe state transitions from the environment. For example, in the problem of plasma control for nuclear fusion, computing the next state for a given state-action pair requires querying an expensive transition function which can lead to many hours of computer simulation or dollars of scientific research. Such expensive data collection prohibits application of standard RL algorithms which usually require a large number of observations to learn. In this work, we address the problem of efficiently learning a policy while making a minimal number of state-action queries to the transition function. In particular, we leverage ideas from Bayesian optimal experimental design to guide the selection of state-action queries for efficient learning. We propose an \emph{acquisition function} that quantifies how much information a state-action pair would provide about the optimal solution to a Markov decision process. At each iteration, our algorithm maximizes this acquisition function, to choose the most informative state-action pair to be queried, thus yielding a data-efficient RL approach. We experiment with a variety of simulated continuous control problems and show that our approach learns an optimal policy with up to 5 -- $1,000\times$ less data than model-based RL baselines and 10^3 -- $10^5\times$ less data than model-free RL baselines. We also provide several ablated comparisons which point to substantial improvements arising from the principled method of obtaining data.

BAM: Bayes with Adaptive Memory

- Josue Nassar, Jennifer Rogers Brennan, Ben Evans, Kendall Lowrey
- abstract@open-review(Poster): Online learning via Bayes' theorem allows new data to be continuously integrated into an agent's current beliefs. However, a naive application of Bayesian methods in non-stationary environments leads to slow adaptation and results in state estimates that may converge confidently to the wrong parameter value. A common solution when learning in changing environments is to discard/downweight past data; however, this simple mechanism of "forgetting" fails to account for the fact that many real-world environments involve revisiting similar states. We propose a new framework, Bayes with Adaptive Memory (BAM), that takes advantage of past experience by allowing the agent to choose which past observations to remember and which to forget. We demonstrate that BAM generalizes many popular Bayesian update rules for non-stationary environments. Through a variety of experiments, we demonstrate the ability of BAM to continuously adapt in an ever-changing world.

Unsupervised Learning of Full-Waveform Inversion: Connecting CNN and Partial Differential Equation in a Loop

- Peng Jin, Xitong Zhang, Yinpeng Chen, Sharon X Huang, Zicheng Liu, Youzuo Lin
- abstract@open-review(Poster): This paper investigates unsupervised learning of Full-Waveform Inversion (FWI), which has been widely used in geophysics to estimate subsurface velocity maps from seismic data. This problem is mathematically formulated by a second order partial differential equation (PDE), but is hard to solve. Moreover, acquiring velocity map is extremely expensive, making it impractical to scale up a supervised approach to train the mapping from seismic data to velocity maps with convolutional neural networks (CNN).We address these difficulties by $\textit{integrating PDE and CNN in a loop}$, thus shifting the paradigm to unsupervised learning that only requires seismic data. In particular, we use finite difference to approximate the forward modeling of PDE as a differentiable operator (from velocity map to seismic data) and model its inversion by CNN (from seismic data to velocity map). Hence, we transform the supervised inversion task into an unsupervised seismic data reconstruction task. We also introduce a new large-scale dataset $\textit{OpenFWI}$, to establish a more challenging benchmark for the community. Experiment results show that our model (using seismic data alone) yields comparable accuracy to the supervised counterpart (using both seismic data and velocity map). Furthermore, it outperforms the supervised model when involving more seismic data.

Conditional Contrastive Learning with Kernel

- Yao-Hung Hubert Tsai, Tianqin Li, Martin Q. Ma, Han Zhao, Kun Zhang, Louis-Philippe Morency, Ruslan Salakhutdinov
- abstract@open-review(Poster): Conditional contrastive learning frameworks consider the conditional sampling procedure that constructs positive or negative data pairs conditioned on specific variables. Fair contrastive learning constructs negative pairs, for example, from the same gender (conditioning on sensitive information), which in turn reduces undesirable information from the learned representations; weakly supervised contrastive learning constructs positive pairs with similar annotative attributes (conditioning on auxiliary information), which in turn are incorporated into the representations. Although conditional contrastive learning enables many applications, the conditional sampling procedure can be challenging if we cannot obtain sufficient data pairs for some values of the conditioning variable. This paper presents Conditional Contrastive Learning with Kernel (CCL-K) that converts existing conditional contrastive objectives into alternative forms that mitigate the insufficient data problem. Instead of sampling data according to the value of the conditioning variable, CCL-K uses the Kernel Conditional Embedding Operator that samples data from all available data and assigns weights to each sampled data given the kernel similarity between the values of the conditioning variable. We conduct experiments using weakly supervised, fair, and hard negatives contrastive learning, showing CCL-K outperforms state-of-the-art baselines.

ConFeSS: A Framework for Single Source Cross-Domain Few-Shot Learning

- Debasmit Das, Sungrack Yun, Fatih Porikli
- abstract@open-review(Poster): Most current few-shot learning methods train a model from abundantly labeled base category data and then transfer and adapt the model to sparsely labeled novel category data. These methods mostly generalize well on novel categories from the same domain as the base categories but perform poorly for distant domain categories. In this paper, we propose a framework for few-shot learning coined as ConFeSS (Contrastive Learning and Feature Selection System) that tackles large domain shift between base and novel categories. The first step of our framework trains a feature extracting backbone with the contrastive loss on the base category data. Since the contrastive loss does not use supervision, the features can generalize better to distant target domains. For the second step, we train a masking module to select relevant features that are more suited to target domain classification. Finally, a classifier is fine-tuned along with the backbone such that the backbone produces features similar to the relevant ones. To evaluate our framework, we tested it on a recently introduced cross-domain few-shot learning benchmark. Experimental results demonstrate that our framework outperforms all meta-learning approaches and produces competitive results against recent cross-domain methods. Additional analyses are also performed to better understand our framework.

Granger causal inference on DAGs identifies genomic loci regulating transcription

- Alexander P Wu, Rohit Singh, Bonnie Berger
- abstract@open-review(Poster): When a dynamical system can be modeled as a sequence of observations, Granger causality is a powerful approach for detecting predictive interactions between its variables. However, traditional Granger causal inference has limited utility in domains where the dynamics need to be represented as directed acyclic graphs (DAGs) rather than as a linear sequence, such as with cell differentiation trajectories. Here, we present GrID-Net, a framework based on graph neural networks with lagged message passing for Granger causal inference on DAG-structured systems. Our motivating application is the analysis of single-cell multimodal data to identify genomic loci that mediate the regulation of specific genes. To our knowledge, GrID-Net is the first single-cell analysis tool that accounts for the temporal lag between a genomic locus becoming accessible and its downstream effect on a target gene's expression. We applied GrID-Net on multimodal single-cell assays that profile chromatin accessibility (ATAC-seq) and gene expression (RNA-seq) in the same cell and show that it dramatically outperforms existing methods for inferring regulatory locus-gene links,

achieving up to 71% greater agreement with independent population genetics-based estimates. By extending Granger causality to DAG-structured dynamical systems, our work unlocks new domains for causal analyses and, more specifically, opens a path towards elucidating gene regulatory interactions relevant to cellular differentiation and complex human diseases at unprecedented scale and resolution.

Energy-Inspired Molecular Conformation Optimization

- Jiaqi Guan, Wesley Wei Qian, qiang liu, Wei-Ying Ma, Jianzhu Ma, Jian Peng
- abstract@open-review(Poster): This paper studies an important problem in computational chemistry: predicting a molecule's spatial atom arrangements, or a molecular conformation. We propose a neural energy minimization formulation that casts the prediction problem into an unrolled optimization process, where a neural network is parametrized to learn the gradient fields of an implicit conformational energy landscape. Assuming different forms of the underlying potential energy function, we can not only reinterpret and unify many of the existing models but also derive new variants of SE(3)-equivariant neural networks in a principled manner. In our experiments, these new variants show superior performance in molecular conformation optimization comparing to existing SE(3)-equivariant neural networks. Moreover, our energy-inspired formulation is also suitable for molecular conformation generation, where we can generate more diverse and accurate conformers comparing to existing baselines.

Towards Deepening Graph Neural Networks: A GNTK-based Optimization Perspective

- Wei Huang, Yayong Li, weitao Du, Richard Xu, Jie Yin, Ling Chen, Miao Zhang
- abstract@open-review(Poster): Graph convolutional networks (GCNs) and their variants have achieved great success in dealing with graph-structured data. Nevertheless, it is well known that deep GCNs suffer from the over-smoothing problem, where node representations tend to be indistinguishable as more layers are stacked up. The theoretical research to date on deep GCNs has focused primarily on expressive power rather than trainability, an optimization perspective. Compared to expressivity, trainability attempts to address a more fundamental question: Given a sufficiently expressive space of models, can we successfully find a good solution via gradient descent-based optimizers? This work fills this gap by exploiting the Graph Neural Tangent Kernel (GNTK), which governs the optimization trajectory under gradient descent for wide GCNs. We formulate the asymptotic behaviors of GNTK in the large depth, which enables us to reveal the dropping trainability of wide and deep GCNs at an exponential rate in the optimization process. Additionally, we extend our theoretical framework to analyze residual connection-based techniques, which are found to be merely able to mitigate the exponential decay of trainability mildly. Inspired by our theoretical insights on trainability, we propose Critical DropEdge, a connectivity-aware and graph-adaptive sampling method, to alleviate the exponential decay problem more fundamentally. Experimental evaluation consistently confirms using our proposed method can achieve better results compared to relevant counterparts with both infinite-width and finite-width.

Connectome-constrained Latent Variable Model of Whole-Brain Neural Activity

- Lu Mi, Richard Xu, Sridhama Prakhya, Albert Lin, Nir Shavit, Aravinthan Samuel, Srinivas C Turaga
- abstract@open-review(Poster): The availability of both anatomical connectivity and brain-wide neural activity measurements in C. elegans make the worm a promising system for learning detailed, mechanistic models of an entire nervous system in a data-driven way. However, one faces several challenges when constructing such a model. We often do not have direct experimental access to important modeling details such as single-neuron dynamics and the signs and strengths of the synaptic connectivity. Further, neural activity can only be measured in a subset of neurons, often indirectly via calcium imaging, and significant trial-to-trial variability has been observed. To address these challenges, we introduce a connectome-constrained latent variable model (CC-LVM) of the unobserved voltage dynamics of the entire C. elegans nervous system and the observed calcium signals. We used the framework of variational autoencoders to fit parameters of the mechanistic simulation constituting the generative model of the LVM to calcium imaging observations. A variational approximate posterior distribution over latent voltage traces for all neurons is efficiently inferred using an inference network, and constrained by a prior distribution given by the biophysical simulation of neural dynamics. We applied this model to an experimental whole-brain dataset, and found that connectomic constraints enable our LVM to predict the activity of neurons whose activity were withheld significantly better than models unconstrained by a connectome. We explored models with different degrees of biophysical detail, and found that models with realistic conductance-based synapses provide markedly better predictions than current-based synapses for this system.

T-WaveNet: A Tree-Structured Wavelet Neural Network for Time Series Signal Analysis

- Minhao LIU, Ailing Zeng, Qiuxia LAI, Ruiyuan Gao, Min Li, Jing Qin, Qiang Xu
- abstract@open-review(Poster): Time series signal analysis plays an essential role in many applications, e.g., activity recognition and healthcare monitoring. Recently, features extracted with deep neural networks (DNNs) have shown to be more effective than conventional hand-crafted ones. However, most existing solutions rely solely on the network to extract information carried in the raw signal, regardless of its inherent physical and statistical properties, leading to sub-optimal performance particularly under a limited amount of training data. In this work, we propose a novel tree-structured wavelet neural network for time series signal analysis, namely \emph{T-WaveNet}, taking advantage of an inherent property of various types of signals, known as the \emph{dominant frequency range}. Specifically, with \emph{T-WaveNet}, we first conduct frequency spectrum energy analysis of the signals to get a set of dominant frequency subbands. Then, we construct a tree-structured network that iteratively decomposes the input signal into various frequency subbands with similar energies. Each node on the tree is built with an invertible neural network (INN) based wavelet transform unit. Such a disentangled representation learning method facilitates a more effective extraction of the discriminative features, as demonstrated with the comprehensive experiments on various real-life time series classification datasets.

Trans-Encoder: Unsupervised sentence-pair modelling through self- and mutual-distillations

- Fangyu Liu, Yunlong Jiao, Jordan Massiah, Emine Yilmaz, Serhii Havrylov
- abstract@open-review(Poster): In NLP, a large volume of tasks involve pairwise comparison between two sequences (e.g. sentence similarity and paraphrase identification). Predominantly, two formulations are used for sentence-pair tasks: bi-encoders and cross-encoders. Bi-encoders produce fixed-dimensional sentence representations and are computationally efficient, however, they usually underperform cross-encoders. Cross-encoders can leverage their attention heads to exploit inter-sentence interactions for better performance but they require task fine-tuning and are computationally more expensive. In this paper, we present a completely unsupervised sentence representation model termed as Trans-Encoder that combines the two learning paradigms into an iterative joint framework to simultaneously learn enhanced bi- and cross-encoders. Specifically, on top of a pre-trained Language Model (PLM), we start with converting it to an unsupervised bi-encoder, and then alternate between the bi- and cross-encoder task formulations. In each alternation, one task formulation will produce pseudo-labels which are used as learning signals for the other task formulation. We then propose an extension to conduct such self-distillation approach on multiple PLMs in parallel and use the average of their pseudo-labels for mutual distillation. Trans-Encoder creates, to the best of our knowledge, the first completely unsupervised cross-encoder and also a state-of-the-art unsupervised bi-encoder for sentence similarity. Both the bi-encoder and cross-encoder formulations of Trans-Encoder outperform recently proposed state-of-the-art unsupervised sentence encoders such as Mirror-BERT and SimCSE by up to 5% on the sentence similarity benchmarks.

Path Integral Sampler: A Stochastic Control Approach For Sampling

- Qinsheng Zhang, Yongxin Chen
- abstract@open-review(Poster): We present Path Integral Sampler~(PIS), a novel algorithm to draw samples from unnormalized probability density functions. The PIS is built on the Schr\"odinger bridge problem which aims to recover the most likely evolution of a diffusion process given its initial distribution and terminal distribution. The PIS draws samples from the initial distribution and then propagates the samples through the Schr\"odinger

bridge to reach the terminal distribution. Applying the Girsanov theorem, with a simple prior diffusion, we formulate the PIS as a stochastic optimal control problem whose running cost is the control energy and terminal cost is chosen according to the target distribution. By modeling the control as a neural network, we establish a sampling algorithm that can be trained end-to-end. We provide theoretical justification of the sampling quality of PIS in terms of Wasserstein distance when sub-optimal control is used. Moreover, the path integrals theory is used to compute importance weights of the samples to compensate for the bias induced by the sub-optimality of the controller and the time-discretization. We experimentally demonstrate the advantages of PIS compared with other start-of-the-art sampling methods on a variety of tasks.

Model Zoo: A Growing Brain That Learns Continually

- Rahul Ramesh, Pratik Chaudhari
- abstract@open-review(Poster): This paper argues that continual learning methods can benefit by splitting the capacity of the learner across multiple models. We use statistical learning theory and experimental analysis to show how multiple tasks can interact with each other in a non-trivial fashion when a single model is trained on them. The generalization error on a particular task can improve when it is trained with synergistic tasks, but can also deteriorate when trained with competing tasks. This theory motivates our method named Model Zoo which, inspired from the boosting literature, grows an ensemble of small models, each of which is trained during one episode of continual learning. We demonstrate that Model Zoo obtains large gains in accuracy on a wide variety of continual learning benchmark problems.

Predicting Physics in Mesh-reduced Space with Temporal Attention

- XU HAN, Han Gao, Tobias Pfaff, Jian-Xun Wang, Liping Liu
- abstract@open-review(Poster): Auto-regressive sequence models for physics prediction are often restricted to low-dimensional systems, as memory cost increases with both spatial extents and sequence length. On the other hand, graph-based next-step prediction models have recently been very successful in modeling complex high-dimensional physical systems on irregular meshes, but suffer from error accumulation and drift, due to their short temporal attention span. In this paper, we present a method that marries the strengths of both approaches. We use a GNN to locally summarize features and create coarsened, compact mesh representation of the system state, onto which we apply a transformer-style temporal attention module. We use a second GNN to decode these predictions back to a full-sized graph and perform fine-scale updates. Our method outperforms a competitive GNN baseline on three complex fluid dynamics prediction tasks, from sonic shocks to vascular flow. We demonstrate stable rollouts without the need for training noise and show perfectly phase-stable predictions even for very long sequences. More broadly, we believe our approach paves the way to bringing the benefits of attention-based sequence models to solving high-dimensional complex physics tasks.

How unlabeled data improve generalization in self-training? A one-hidden-layer theoretical analysis

- Shuai Zhang, Meng Wang, Sijia Liu, Pin-Yu Chen, Jinjun Xiong
- abstract@open-review(Poster): Self-training, a semi-supervised learning algorithm, leverages a large amount of unlabeled data to improve learning when the labeled data are limited. Despite empirical successes, its theoretical characterization remains elusive. To the best of our knowledge, this work establishes the first theoretical analysis for the known iterative self-training paradigm and formally proves the benefits of unlabeled data in both training convergence and generalization ability. To make our theoretical analysis feasible, we focus on the case of one-hidden-layer neural networks. However, theoretical understanding of iterative self-training is non-trivial even for a shallow neural network. One of the key challenges is that existing neural network landscape analysis built upon supervised learning no longer holds in the (semi-supervised) self-training paradigm. We address this challenge and prove that iterative self-training converges linearly with both convergence rate and generalization accuracy improved in the order of $1/\sqrt{M}$, where M is the number of unlabeled samples. Extensive experiments from shallow neural networks to deep neural networks are also provided to justify the correctness of our established theoretical insights on self-training.

Learning to Dequantise with Truncated Flows

- Shawn Tan, Chin-Wei Huang, Alessandro Sordoni, Aaron Courville
- abstract@open-review(Poster): Dequantisation is a general technique used for transforming data described by a discrete random variable x into a continuous (latent) random variable z, for the purpose of it being modeled by likelihood-based density models. Dequantisation was first introduced in the context of ordinal data, such as image pixel values. However, when the data is categorical, the dequantisation scheme is not obvious. We learn such a dequantisation scheme $q(z | x)$, using variational inference with TRUncated FLows (TRUFL) --- a novel flow-based model that allows the dequantiser to have a learnable truncated support. Unlike previous work, the TRUFL dequantiser is (i) capable of embedding the data losslessly in certain cases, since the truncation allows the conditional distributions $q(z | x)$ to have non-overlapping bounded supports, while being (ii) trainable with back-propagation. Addtionally, since the support of the marginal $q(z)$ is bounded and the support of prior $p(z)$ is not, we propose renormalising the prior distribution over the support of $q(z)$. We derive a lower bound for training, and propose a rejection sampling scheme to account for the invalid samples during generation. Experimentally, we benchmark TRUFL on constrained generation tasks, and find that it outperforms prior approaches. In addition, we find that rejection sampling results in higher validity for the constrained problems.

Curriculum learning as a tool to uncover learning principles in the brain

- Daniel R. Kepple, Rainer Engelken, Kanaka Rajan
- abstract@open-review(Poster): We present a novel approach to use curricula to identify principles by which a system learns. Previous work in curriculum learning has focused on how curricula can be designed to improve learning of a model on particular tasks. We consider the inverse problem: what can a curriculum tell us about how a learning system acquired a task? Using recurrent neural networks (RNNs) and models of common experimental neuroscience tasks, we demonstrate that curricula can be used to differentiate learning principles using target-based and a representation-based loss functions as use cases. In particular, we compare the performance of RNNs using target-based learning rules versus those using representational learning rules on three different curricula in the context of two tasks. We show that the learned state-space trajectories of RNNs trained by these two learning rules under all curricula tested are indistinguishable. However, by comparing learning times during different curricula, we can disambiguate the learning rules and challenge traditional approaches of interrogating learning systems. Although all animals in neuroscience lab settings are trained by curriculum-based procedures called shaping, almost no behavioral or neural data are collected or published on the relative successes or training times under different curricula. Our results motivate the systematic collection and curation of data during shaping by demonstrating curriculum learning in RNNs as a tool to probe and differentiate learning principles used by biological systems, over conventional statistical analyses of learned state spaces.

Optimizer Amalgamation

- Tianshu Huang, Tianlong Chen, Sijia Liu, Shiyu Chang, Lisa Amini, Zhangyang Wang
- abstract@open-review(Poster): Selecting an appropriate optimizer for a given problem is of major interest for researchers and practitioners. Many analytical optimizers have been proposed using a variety of theoretical and empirical approaches; however, none can offer a universal advantage over other competitive optimizers. We are thus motivated to study a new problem named Optimizer Amalgamation: how can we best combine a pool of "teacher" optimizers into a single "student" optimizer that can have stronger problem-specific performance? In this paper, we draw inspiration from the field of "learning to optimize" to use a learnable amalgamation target. First, we define three differentiable amalgamation mechanisms to amalgamate a pool of analytical optimizers by gradient descent. Then, in order to reduce variance of the amalgamation process, we also explore methods to stabilize the

amalgamation process by perturbing the amalgamation target. Finally, we present experiments showing the superiority of our amalgamated optimizer compared to its amalgamated components and learning to optimize baselines, and the efficacy of our variance reducing perturbations.

An Agnostic Approach to Federated Learning with Class Imbalance

- Zebang Shen, Juan Cervino, Hamed Hassani, Alejandro Ribeiro
- abstract@open-review(Poster): Federated Learning (FL) has emerged as the tool of choice for training deep models over heterogeneous and decentralized datasets. As a reflection of the experiences from different clients, severe class imbalance issues are observed in real-world FL problems. Moreover, there exists a drastic mismatch between the imbalances from the local and global perspectives, i.e. a local majority class can be the minority of the population. Additionally, the privacy requirement of FL poses an extra challenge, as one should handle class imbalance without identifying the minority class. In this paper we propose a novel agnostic constrained learning formulation to tackle the class imbalance problem in FL, without requiring further information beyond the standard FL objective. A meta algorithm, CLIMB, is designed to solve the target optimization problem, with its convergence property analyzed under certain oracle assumptions. Through an extensive empirical study over various data heterogeneity and class imbalance configurations, we showcase that CLIMB considerably improves the performance in the minority class without compromising the overall accuracy of the classifier, which significantly outperforms previous arts. In fact, we observe the greatest performance boost in the most difficult scenario where every client only holds data from one class. The code can be found here https://github.com/shenzebang/Federated-Learning-Pytorch.

A Fine-Tuning Approach to Belief State Modeling

- Samuel Sokota, Hengyuan Hu, David J Wu, J Zico Kolter, Jakob Nicolaus Foerster, Noam Brown
- abstract@open-review(Poster): We investigate the challenge of modeling the belief state of a partially observable Markov system, given sample-access to its dynamics model. This problem setting is often approached using parametric sequential generative modeling methods. However, these methods do not leverage any additional computation at inference time to increase their accuracy. Moreover, applying these methods to belief state modeling in certain multi-agent settings would require passing policies into the belief model---at the time of writing, there have been no successful demonstrations of this. Toward addressing these shortcomings, we propose an inference-time improvement framework for parametric sequential generative modeling methods called belief fine-tuning (BFT). BFT leverages approximate dynamic programming in the form of fine-tuning to determine the model parameters at each time step. It can improve the accuracy of the belief model at test time because it specializes the model to the space of local observations. Furthermore, because this specialization occurs after the action or policy has already been decided, BFT does not require the belief model to process it as input. As a result of the latter point, BFT enables, for the first time, approximate public belief state search in imperfect-information games where the number of possible information states is too large to track tabularly. We exhibit these findings on large-scale variants of the benchmark game Hanabi.

Differentially Private Fine-tuning of Language Models

- Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, Huishuai Zhang
- abstract@open-review(Poster): We give simpler, sparser, and faster algorithms for differentially private fine-tuning of large-scale pre-trained language models, which achieve the state-of-the-art privacy versus utility tradeoffs on many standard NLP tasks. We propose a meta-framework for this problem, inspired by the recent success of highly parameter-efficient methods for fine-tuning. Our experiments show that differentially private adaptations of these approaches outperform previous private algorithms in three important dimensions: utility, privacy, and the computational and memory cost of private training. On many commonly studied datasets, the utility of private models approaches that of non-private models. For example, on the MNLI dataset we achieve an accuracy of 87.8% using RoBERTa-Large and 83.5% using RoBERTa-Base with a privacy budget of $\epsilon = 6.7$. In comparison, absent privacy constraints, RoBERTa-Large achieves an accuracy of 90.2%. Our findings are similar for natural language generation when privately fine-tuning GPT-2. Our experiments also show that larger models are better suited for private fine-tuning: while they are well known to achieve superior accuracy non-privately, we find that they also better maintain their accuracy when privacy is introduced.

P-Adapters: Robustly Extracting Factual Information from Language Models with Diverse Prompts

- Benjamin Newman, Prafulla Kumar Choubey, Nazneen Rajani
- abstract@open-review(Poster): Recent work (e.g. LAMA (Petroni et al., 2019)) has found that the quality of the factual information extracted from Large Language Models (LLMs) depends on the prompts used to query them. This inconsistency is problematic because different users will query LLMs for the same information using different wording, but should receive the same, accurate responses regardless. In this work we aim to address this shortcoming by introducing P-Adapters: lightweight models that sit between the embedding layer and first attention layer of LLMs. They take LLM embeddings as input and output continuous prompts that are used to query the LLM. Additionally, we investigate Mixture of Experts (MoE) models that learn a set of continuous prompts (the "experts") and select one to query the LLM. These require a separate classifier trained on human-annotated data to map natural language prompts to the continuous ones. P-Adapters perform comparably to the more complex MoE models in extracting factual information from BERT and RoBERTa while eliminating the need for additional annotations. P-Adapters show between 12-26% absolute improvement in precision and 36-50% absolute improvement in consistency over a baseline of just using natural language queries alone. Finally, we investigate what makes P-Adapters successful and conclude that a significant factor is access to the LLM's embeddings of the original natural language prompt, particularly the subject of the entity pair being queried.

Iterated Reasoning with Mutual Information in Cooperative and Byzantine Decentralized Teaming

- Sachin G Konan, Esmaeil Seraj, Matthew Gombolay
- abstract@open-review(Poster): Information sharing is key in building team cognition and enables coordination and cooperation. High-performing human teams also benefit from acting strategically with hierarchical levels of iterated communication and rationalizability, meaning a human agent can reason about the actions of their teammates in their decision-making. Yet, the majority of prior work in Multi-Agent Reinforcement Learning (MARL) does not support iterated rationalizability and only encourage inter-agent communication, resulting in a suboptimal equilibrium cooperation strategy. In this work, we show that reformulating an agent's policy to be conditional on the policies of its neighboring teammates inherently maximizes Mutual Information (MI) lower-bound when optimizing under Policy Gradient (PG). Building on the idea of decision-making under bounded rationality and cognitive hierarchy theory, we show that our modified PG approach not only maximizes local agent rewards but also implicitly reasons about MI between agents without the need for any explicit ad-hoc regularization terms. Our approach, InfoPG, outperforms baselines in learning emergent collaborative behaviors and sets the state-of-the-art in decentralized cooperative MARL tasks. Our experiments validate the utility of InfoPG by achieving higher sample efficiency and significantly larger cumulative reward in several complex cooperative multi-agent domains.

Step-unrolled Denoising Autoencoders for Text Generation

- Nikolay Savinov, Junyoung Chung, Mikolaj Binkowski, Erich Elsen, Aaron van den Oord
- abstract@open-review(Poster): In this paper we propose a new generative model of text, Step-unrolled Denoising Autoencoder (SUNDAE), that does not rely on autoregressive models. Similarly to denoising diffusion techniques, SUNDAE is repeatedly applied on a sequence of tokens, starting from random inputs and improving them each time until convergence. We present a simple new improvement operator that converges in fewer iterations than diffusion methods, while qualitatively producing better samples on natural language datasets. SUNDAE achieves state-of-the-art results (among non-autoregressive methods) on the WMT'14 English-to-German translation task and good qualitative results on unconditional language modeling on the Colossal Cleaned

Common Crawl dataset and a dataset of Python code from GitHub. The non-autoregressive nature of SUNDAE opens up possibilities beyond left-to-right prompted generation, by filling in arbitrary blank patterns in a template.

Hindsight Foresight Relabeling for Meta-Reinforcement Learning

- Michael Wan, Jian Peng, Tanmay Gangwani
- abstract@open-review(Poster): Meta-reinforcement learning (meta-RL) algorithms allow for agents to learn new behaviors from small amounts of experience, mitigating the sample inefficiency problem in RL. However, while meta-RL agents can adapt quickly to new tasks at test time after experiencing only a few trajectories, the meta-training process is still sample-inefficient. Prior works have found that in the multi-task RL setting, relabeling past transitions and thus sharing experience among tasks can improve sample efficiency and asymptotic performance. We apply this idea to the meta-RL setting and devise a new relabeling method called Hindsight Foresight Relabeling (HFR). We construct a relabeling distribution using the combination of "hindsight", which is used to relabel trajectories using reward functions from the training task distribution, and "foresight", which takes the relabeled trajectories and computes the utility of each trajectory for each task. HFR is easy to implement and readily compatible with existing meta-RL algorithms. We find that HFR improves performance when compared to other relabeling methods on a variety of meta-RL tasks.

LoRA: Low-Rank Adaptation of Large Language Models

- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen
- abstract@open-review(Poster): An important paradigm of natural language processing consists of large-scale pre-training on general domain data and adaptation to particular tasks or domains. As we pre-train larger models, full fine-tuning, which retrains all model parameters, becomes less feasible. Using GPT-3 175B as an example -- deploying independent instances of fine-tuned models, each with 175B parameters, is prohibitively expensive. We propose Low-Rank Adaptation, or LoRA, which freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks. Compared to GPT-3 175B fine-tuned with Adam, LoRA can reduce the number of trainable parameters by a factor of 10,000 and the GPU memory requirement by a factor of 3. LoRA performs on-par or better than fine-tuning in model quality on RoBERTa, DeBERTa, GPT-2, and GPT-3, despite having fewer trainable parameters, a higher training throughput, and, unlike adapters, no additional inference latency. We also provide an empirical investigation into rank-deficiency in language model adaptation, which sheds light on the efficacy of LoRA. We release a package that facilitates the integration of LoRA with PyTorch models and provide our implementations and model checkpoints for RoBERTa, DeBERTa, and GPT-2 at https://github.com/microsoft/LoRA.

Which Shortcut Cues Will DNNs Choose? A Study from the Parameter-Space Perspective

- Luca Scimeca, Seong Joon Oh, Sanghyuk Chun, Michael Poli, Sangdoo Yun
- abstract@open-review(Poster): Deep neural networks (DNNs) often rely on easyâ€“toâ€“learn discriminatory features, or cues, that are not necessarily essential to the problem at hand. For example, ducks in an image may be recognized based on their typical background scenery, such as lakes or streams. This phenomenon, also known as shortcut learning, is emerging as a key limitation of the current generation of machine learning models. In this work, we introduce a set of experiments to deepen our understanding of shortcut learning and its implications. We design a training setup with several shortcut cues, named WCST-ML, where each cue is equally conducive to the visual recognition problem at hand. Even under equal opportunities, we observe that (1) certain cues are preferred to others, (2) solutions biased to the easyâ€“toâ€“learn cues tend to converge to relatively flat minima on the loss surface, and (3) the solutions focusing on those preferred cues are far more abundant in the parameter space. We explain the abundance of certain cues via their Kolmogorov (descriptional) complexity: solutions corresponding to Kolmogorov-simple cues are abundant in the parameter space and are thus preferred by DNNs. Our studies are based on the synthetic dataset DSprites and the face dataset UTKFace. In our WCST-ML, we observe that the inborn bias of models leans toward simple cues, such as color and ethnicity. Our findings emphasize the importance of active human intervention to remove the inborn model biases that may cause negative societal impacts.

Efficient Computation of Deep Nonlinear Infinite-Width Neural Networks that Learn Features

- Greg Yang, Michael Santacroce, Edward J Hu
- abstract@open-review(Poster): While a popular limit of infinite-width neural networks, the Neural Tangent Kernel (NTK) often exhibits performance gaps from finite-width neural networks on standard datasets, due to lack of feature learning. Although the feature learning *maximal update limit*, or *Î¼-limit* (Yang and Hu, 2020) of wide networks has closed the gap for 1-hidden-layer linear models, no one has been able to demonstrate this for deep nonlinear multi-layer perceptrons (MLP) because of Î¼-limitâ€™s computational difficulty in this setting. Here, we solve this problem by proposing a novel feature learning limit, the *Ï€-limit*, that bypasses the computational issues. The Ï€-limit, in short, is the limit of a form of projected gradient descent, and the Ï€-limit of an MLP is roughly another MLP where gradients are appended to weights during training. We prove its almost sure convergence with width using the Tensor Programs technique. We evaluate it on CIFAR10 and Omniglot against NTK as well as finite networks, finding the Ï€-limit outperform finite-width models trained normally (without projection) in both settings, closing the performance gap between finite- and infinite-width neural networks previously left by NTK. Code for this work is available at github.com/santacml/pilim.

TRAIL: Near-Optimal Imitation Learning with Suboptimal Data

- Mengjiao Yang, Sergey Levine, Ofir Nachum
- abstract@open-review(Poster): In imitation learning, one aims to learn task-solving policies using access to near-optimal expert trajectories collected from the task environment. However, high-quality trajectories -- e.g., from human experts -- can be expensive to obtain in practical settings. On the contrary, it is often much easier to obtain large amounts of suboptimal trajectories which can nevertheless provide insight into the structure of the environment, showing what \emph{could} be done in the environment even if not what \emph{should} be done. Is it possible to formalize these conceptual benefits and devise algorithms to use offline datasets to yield \emph{provable} improvements to the sample-efficiency of imitation learning? In this work, we answer this question affirmatively and present training objectives which use an offline dataset to learn an approximate \emph{factored} dynamics model whose structure enables the extraction of a \emph{latent action space}. Our theoretical analysis shows that the learned latent action space can boost the sample-efficiency of downstream imitation learning, effectively reducing the need for large near-optimal expert datasets through the use of auxiliary non-expert data. We evaluate the practicality of our objective through experiments on a set of navigation and locomotion tasks. Our results verify the benefits suggested by our theory and show that our algorithms is able to recover near-optimal policies with fewer expert trajectories.

On the benefits of maximum likelihood estimation for Regression and Forecasting

- Pranjal Awasthi, Abhimanyu Das, Rajat Sen, Ananda Theertha Suresh
- abstract@open-review(Poster): We advocate for a practical Maximum Likelihood Estimation (MLE) approach towards designing loss functions for regression and forecasting, as an alternative to the typical approach of direct empirical risk minimization on a specific target metric. The MLE approach is better suited to capture inductive biases such as prior domain knowledge in datasets, and can output post-hoc estimators at inference time that can optimize different types of target metrics. We present theoretical results to demonstrate that our approach is competitive with any estimator for the target metric under some general conditions. In two example practical settings, Poisson and Pareto regression, we show that our competitive results can be used to prove that the MLE approach has better excess risk bounds than directly minimizing the target metric. We also demonstrate empirically that our method instantiated with a well-designed general purpose mixture likelihood family can obtain superior performance for a variety of tasks across time-series forecasting and regression datasets with different data distributions.

Effect of scale on catastrophic forgetting in neural networks

- Vinay Venkatesh Ramasesh, Aitor Lewkowycz, Ethan Dyer
- abstract@open-review(Poster): Catastrophic forgetting presents a challenge in developing deep learning models capable of continual learning, i.e. learning tasks sequentially. Recently, both computer vision and natural-language processing have witnessed great progress through the use of large-scale pretrained models. In this work, we present an empirical study of catastrophic forgetting in this pretraining paradigm. Our experiments indicate that large, pretrained ResNets and Transformers are significantly more resistant to forgetting than randomly-initialized, trained-from-scratch models; this robustness systematically improves with scale of both model and pretraining dataset size. We take initial steps towards characterizing what aspect of model representations allows them to perform continual learning so well, finding that in the pretrained models, distinct class representations grow more orthogonal with scale. Our results suggest that, when possible, scale and a diverse pretraining dataset can be useful ingredients in mitigating catastrophic forgetting.

Learn Locally, Correct Globally: A Distributed Algorithm for Training Graph Neural Networks

- Morteza Ramezani, Weilin Cong, Mehrdad Mahdavi, Mahmut Kandemir, Anand Sivasubramaniam
- abstract@open-review(Poster): Despite the recent success of Graph Neural Networks (GNNs), training GNNs on large graphs remains challenging. The limited resource capacities of the existing servers, the dependency between nodes in a graph, and the privacy concern due to the centralized storage and model learning have spurred the need to design an effective distributed algorithm for GNN training. However, existing distributed GNN training methods impose either excessive communication costs or large memory overheads that hinders their scalability. To overcome these issues, we propose a communication-efficient distributed GNN training technique named $\text{\textit{Learn Locally, Correct Globally}}$ (LLCG). To reduce the communication and memory overhead, each local machine in LLCG first trains a GNN on its local data by ignoring the dependency between nodes among different machines, then sends the locally trained model to the server for periodic model averaging. However, ignoring node dependency could result in significant performance degradation. To solve the performance degradation, we propose to apply $\text{\textit{Global Server Corrections}}$ on the server to refine the locally learned models. We rigorously analyze the convergence of distributed methods with periodic model averaging for training GNNs and show that naively applying periodic model averaging but ignoring the dependency between nodes will suffer from an irreducible residual error. However, this residual error can be eliminated by utilizing the proposed global corrections to entail fast convergence rate. Extensive experiments on real-world datasets show that LLCG can significantly improve the efficiency without hurting the performance.

Conditional Image Generation by Conditioning Variational Auto-Encoders

- William Harvey, Saeid Naderiparizi, Frank Wood
- abstract@open-review(Poster): We present a conditional variational auto-encoder (VAE) which, to avoid the substantial cost of training from scratch, uses an architecture and training objective capable of leveraging a foundation model in the form of a pretrained unconditional VAE. To train the conditional VAE, we only need to train an artifact to perform amortized inference over the unconditional VAE's latent variables given a conditioning input. We demonstrate our approach on tasks including image inpainting, for which it outperforms state-of-the-art GAN-based approaches at faithfully representing the inherent uncertainty. We conclude by describing a possible application of our inpainting model, in which it is used to perform Bayesian experimental design for the purpose of guiding a sensor.

Learning 3D Representations of Molecular Chirality with Invariance to Bond Rotations

- Keir Adams, Lagnajit Pattanaik, Connor W. Coley
- abstract@open-review(Poster): Molecular chirality, a form of stereochemistry most often describing relative spatial arrangements of bonded neighbors around tetrahedral carbon centers, influences the set of 3D conformers accessible to the molecule without changing its 2D graph connectivity. Chirality can strongly alter (bio)chemical interactions, particularly protein-drug binding. Most 2D graph neural networks (GNNs) designed for molecular property prediction at best use atomic labels to naÃ¯vely treat chirality, while E(3)-invariant 3D GNNs are invariant to chirality altogether. To enable representation learning on molecules with defined stereochemistry, we design an SE(3)-invariant model that processes torsion angles of a 3D molecular conformer. We explicitly model conformational flexibility by integrating a novel type of invariance to rotations about internal molecular bonds into the architecture, mitigating the need for multi-conformer data augmentation. We test our model on four benchmarks: contrastive learning to distinguish conformers of different stereoisomers in a learned latent space, classification of chiral centers as R/S, prediction of how enantiomers rotate circularly polarized light, and ranking enantiomers by their docking scores in an enantiosensitive protein pocket. We compare our model, Chiral InterRoto-Invariant Neural Network (ChIRo), with 2D and 3D GNNs to demonstrate that our model achieves state of the art performance when learning chiral-sensitive functions from molecular structures.

Neural Methods for Logical Reasoning over Knowledge Graphs

- Alfonso Amayuelas, Shuai Zhang, Xi Susie Rao, Ce Zhang
- abstract@open-review(Poster): Reasoning is a fundamental problem for computers and deeply studied in Artificial Intelligence. In this paper, we specifically focus on answering multi-hop logical queries on Knowledge Graphs (KGs). This is a complicated task because, in real world scenarios, the graphs tend to be large and incomplete. Most previous works have been unable to create models that accept full First-Order Logical (FOL) queries, which includes negative queries, and have only been able to process a limited set of query structures. Additionally, most methods present logic operators that can only perform the logical operation they are made for. We introduce a set of models that use Neural Networks to create one-point vector embeddings to answer the queries. The versatility of neural networks allows the framework to handle FOL queries with Conjunction, Disjunction and Negation operators. We demonstrate experimentally the performance of our models through extensive experimentation on well-known benchmarking datasets. Besides having more versatile operators, the models achieve a 10% relative increase over best performing state of the art and more than 30% over the original method based on single-point vector embeddings.

Consistent Counterfactuals for Deep Models

- Emily Black, Zifan Wang, Matt Fredrikson
- abstract@open-review(Poster): Counterfactual examples are one of the most commonly-cited methods for explaining the predictions of machine learning models in key areas such as finance and medical diagnosis. Counterfactuals are often discussed under the assumption that the model on which they will be used is static, but in deployment models may be periodically retrained or fine-tuned. This paper studies the consistency of model prediction on counterfactual examples in deep networks under small changes to initial training conditions, such as weight initialization and leave-one-out variations in data, as often occurs during model deployment. We demonstrate experimentally that counterfactual examples for deep models are often inconsistent across such small changes, and that increasing the cost of the counterfactual, a stability-enhancing mitigation suggested by prior work in the context of simpler models, is not a reliable heuristic in deep networks. Rather, our analysis shows that a model's Lipschitz continuity around the counterfactual, along with confidence of its prediction, is key to its consistency across related models. To this end, we propose Stable Neighbor Search as a way to generate more consistent counterfactual explanations, and illustrate the effectiveness of this approach on several benchmark datasets.

Unified Visual Transformer Compression

- Shixing Yu, Tianlong Chen, Jiayi Shen, Huan Yuan, Jianchao Tan, Sen Yang, Ji Liu, Zhangyang Wang
- abstract@open-review(Poster): Vision transformers (ViTs) have gained popularity recently. Even without customized image operators such as convolutions, ViTs can yield competitive performance when properly trained on massive data. However, the computational overhead of ViTs remains prohibitive, due to stacking multi-head self-attention modules and else. Compared to the vast literature and prevailing success in compressing convolutional neural networks, the study of Vision Transformer compression has also just emerged, and existing works focused on one or two aspects of compression. This paper proposes a unified ViT compression framework that seamlessly assembles three effective techniques: pruning, layer skipping, and knowledge distillation. We formulate a budget-constrained, end-to-end optimization framework, targeting jointly learning model weights, layer-wise pruning ratios/masks, and skip configurations, under a distillation loss. The optimization problem is then solved using the primal-dual algorithm. Experiments are conducted with several ViT variants, e.g. DeiT and T2T-ViT backbones on the ImageNet dataset, and our approach consistently outperforms recent competitors. For example, DeiT-Tiny can be trimmed down to 50\% of the original FLOPs almost without losing accuracy. Codes are available online:~\url{https://github.com/VITA-Group/UVC}.

Transformer-based Transform Coding

- Yinhao Zhu, Yang Yang, Taco Cohen
- abstract@open-review(Poster): Neural data compression based on nonlinear transform coding has made great progress over the last few years, mainly due to improvements in prior models, quantization methods and nonlinear transforms. A general trend in many recent works pushing the limit of rate-distortion performance is to use ever more expensive prior models that can lead to prohibitively slow decoding. Instead, we focus on more expressive transforms that result in a better rate-distortion-computation trade-off. Specifically, we show that nonlinear transforms built on Swin-transformers can achieve better compression efficiency than transforms built on convolutional neural networks (ConvNets), while requiring fewer parameters and shorter decoding time. Paired with a compute-efficient Channel-wise Auto-Regressive Model prior, our SwinT-ChARM model outperforms VTM-12.1 by 3.68% in BD-rate on Kodak with comparable decoding speed. In P-frame video compression setting, we are able to outperform the popular ConvNet-based scale-space-flow model by 12.35% in BD-rate on UVG. We provide model scaling studies to verify the computational efficiency of the proposed solutions and conduct several analyses to reveal the source of coding gain of transformers over ConvNets, including better spatial decorrelation, flexible effective receptive field, and more localized response of latent pixels during progressive decoding.

Object Pursuit: Building a Space of Objects via Discriminative Weight Generation

- Chuanyu Pan, Yanchao Yang, Kaichun Mo, Yueqi Duan, Leonidas Guibas
- abstract@open-review(Poster): We propose a framework to continuously learn object-centric representations for visual learning and understanding. Existing object-centric representations either rely on supervisions that individualize objects in the scene, or perform unsupervised disentanglement that can hardly deal with complex scenes in the real world. To mitigate the annotation burden and relax the constraints on the statistical complexity of the data, our method leverages interactions to effectively sample diverse variations of an object and the corresponding training signals while learning the object-centric representations. Throughout learning, objects are streamed one by one in random order with unknown identities, and are associated with latent codes that can synthesize discriminative weights for each object through a convolutional hypernetwork. Moreover, re-identification of learned objects and forgetting prevention are employed to make the learning process efficient and robust. We perform an extensive study of the key features of the proposed framework and analyze the characteristics of the learned representations. Furthermore, we demonstrate the capability of the proposed framework in learning representations that can improve label efficiency in downstream tasks. Our code and trained models are made publicly available at: https://github.com/pptrick/Object-Pursuit.

PAC Prediction Sets Under Covariate Shift

- Sangdon Park, Edgar Dobriban, Insup Lee, Osbert Bastani
- abstract@open-review(Poster): An important challenge facing modern machine learning is how to rigorously quantify the uncertainty of model predictions. Conveying uncertainty is especially important when there are changes to the underlying data distribution that might invalidate the predictive model. Yet, most existing uncertainty quantification algorithms break down in the presence of such shifts. We propose a novel approach that addresses this challenge by constructing \emph{probably approximately correct (PAC)} prediction sets in the presence of covariate shift. Our approach focuses on the setting where there is a covariate shift from the source distribution (where we have labeled training examples) to the target distribution (for which we want to quantify uncertainty). Our algorithm assumes given importance weights that encode how the probabilities of the training examples change under the covariate shift. In practice, importance weights typically need to be estimated; thus, we extend our algorithm to the setting where we are given confidence intervals for the importance weights. We demonstrate the effectiveness of our approach on covariate shifts based on DomainNet and ImageNet. Our algorithm satisfies the PAC constraint, and gives prediction sets with the smallest average normalized size among approaches that always satisfy the PAC constraint.

Generalization of Neural Combinatorial Solvers Through the Lens of Adversarial Robustness

- Simon Geisler, Johanna Sommer, Jan Schuchardt, Aleksandar Bojchevski, Stephan GÃ¼nnemann
- abstract@open-review(Poster): End-to-end (geometric) deep learning has seen first successes in approximating the solution of combinatorial optimization problems. However, generating data in the realm of NP-hard/-complete tasks brings practical and theoretical challenges, resulting in evaluation protocols that are too optimistic. Specifically, most datasets only capture a simpler subproblem and likely suffer from spurious features. We investigate these effects by studying adversarial robustness -a local generalization property- to reveal hard, model-specific instances and spurious features. For this purpose, we derive perturbation models for SAT and TSP. Unlike in other applications, where perturbation models are designed around subjective notions of imperceptibility, our perturbation models are efficient and sound, allowing us to determine the true label of perturbed samples without a solver. Surprisingly, with such perturbations, a sufficiently expressive neural solver does not suffer from the limitations of the accuracy-robustness trade-off common in supervised learning. Although such robust solvers exist, we show empirically that the assessed neural solvers do not generalize well w.r.t. small perturbations of the problem instance.

One After Another: Learning Incremental Skills for a Changing World

- Nur Muhammad Mahi Shafiullah, Lerrel Pinto
- abstract@open-review(Poster): Reward-free, unsupervised discovery of skills is an attractive alternative to the bottleneck of hand-designing rewards in environments where task supervision is scarce or expensive. However, current skill pre-training methods, like many RL techniques, make a fundamental assumption -- stationary environments during training. Traditional methods learn all their skills simultaneously, which makes it difficult for them to both quickly adapt to changes in the environment, and to not forget earlier skills after such adaptation. On the other hand, in an evolving or expanding environment, skill learning must be able to adapt fast to new environment situations while not forgetting previously learned skills. These two conditions make it difficult for classic skill discovery to do well in an evolving environment. In this work, we propose a new framework for skill discovery, where skills are learned one after another in an incremental fashion. This framework allows newly learned skills to adapt to new environment or agent dynamics, while the fixed old skills ensure the agent doesn't forget a learned skill. We demonstrate experimentally that in both evolving and static environments, incremental skills significantly outperform current state-of-the-art skill discovery methods on both skill quality and the ability to solve downstream tasks. Videos for learned skills and code are made public on https://notmahi.github.io/disk

Graph-Guided Network for Irregularly Sampled Multivariate Time Series

- Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, Marinka Zitnik
- abstract@open-review(Poster): In many domains, including healthcare, biology, and climate science, time series are irregularly sampled with varying time intervals between successive readouts and different subsets of variables (sensors) observed at different time points. Here, we introduce RAINDROP, a graph neural network that embeds irregularly sampled and multivariate time series while also learning the dynamics of sensors purely from observational data. RAINDROP represents every sample as a separate sensor graph and models time-varying dependencies between sensors with a novel message passing operator. It estimates the latent sensor graph structure and leverages the structure together with nearby observations to predict misaligned readouts. This model can be interpreted as a graph neural network that sends messages over graphs that are optimized for capturing time-varying dependencies among sensors. We use RAINDROP to classify time series and interpret temporal dynamics on three healthcare and human activity datasets. RAINDROP outperforms state-of-the-art methods by up to 11.4% (absolute F1-score points), including techniques that deal with irregular sampling using fixed discretization and set functions. RAINDROP shows superiority in diverse setups, including challenging leave-sensor-out settings.

FILM: Following Instructions in Language with Modular Methods

- So Yeon Min, Devendra Singh Chaplot, Pradeep Kumar Ravikumar, Yonatan Bisk, Ruslan Salakhutdinov
- abstract@open-review(Poster): Recent methods for embodied instruction following are typically trained end-to-end using imitation learning. This often requires the use of expert trajectories and low-level language instructions. Such approaches assume that neural states will integrate multimodal semantics to perform state tracking, building spatial memory, exploration, and long-term planning. In contrast, we propose a modular method with structured representations that (1) builds a semantic map of the scene and (2) performs exploration with a semantic search policy, to achieve the natural language goal. Our modular method achieves SOTA performance (24.46 %) with a substantial (8.17 % absolute) gap from previous work while using less data by eschewing both expert trajectories and low-level instructions. Leveraging low-level language, however, can further increase our performance (26.49 %). Our findings suggest that an explicit spatial memory and a semantic search policy can provide a stronger and more general representation for state-tracking and guidance, even in the absence of expert trajectories or low-level instructions.

The Evolution of Uncertainty of Learning in Games

- Yun Kuen Cheung, Georgios Piliouras, Yixin Tao
- abstract@open-review(Poster): Learning in games has become an object of intense interest for ML due to its connections to numerous AI architectures. We study standard online learning in games but from a non-standard perspective. Instead of studying the behavior of a single initial condition and whether it converges to equilibrium or not, we study the behavior of a probability distribution/measure over a set of initial conditions. This initial uncertainty is well-motivated both from a standard game-theoretic perspective (e.g. a modeler's uncertainty about the agents' initial beliefs) as well as from a ML one (e.g. noisy measurements, system initialization from a dataset distribution). Despite this, little is formally known about whether and under what conditions uncertainty is amplified or reduced in these systems. We use the popular measure of differential entropy to quantify the evolution of uncertainty. We find that such analysis shares an intimate relationship with volume analysis, a technique which was recently used to demonstrate the occurrence of Lyapunov chaos when using Multiplicative Weights Update (MWU) or Follow-the-Regularized-Leader (FTRL) algorithms in zero-sum games. This allows us to show that the differential entropy of these learning-in-game systems increases linearly with time, formalizing their increased unpredictability over time. We showcase the power of the framework by applying it in the study of multiple related systems, including different standard online optimization algorithms in numerous games and dynamics of evolutionary game theory.

Explainable GNN-Based Models over Knowledge Graphs

- David Jaime Tena Cucala, Bernardo Cuenca Grau, Egor V. Kostylev, Boris Motik
- abstract@open-review(Poster): Graph Neural Networks (GNNs) are often used to learn transformations of graph data. While effective in practice, such approaches make predictions via numeric manipulations so their output cannot be easily explained symbolically. We propose a new family of GNN-based transformations of graph data that can be trained effectively, but where all predictions can be explained symbolically as logical inferences in Datalogâ€”a well-known rule-based formalism. In particular, we show how to encode an input knowledge graph into a graph with numeric feature vectors, process this graph using a GNN, and decode the result into an output knowledge graph. We use a new class of monotonic GNNs (MGNNs) to ensure that this process is equivalent to a round of application of a set of Datalog rules. We also show that, given an arbitrary MGNN, we can automatically extract rules that completely characterise the transformation. We evaluate our approach by applying it to classification tasks in knowledge graph completion.

Mention Memory: incorporating textual knowledge into Transformers through entity mention attention

- Michiel de Jong, Yury Zemlyanskiy, Nicholas FitzGerald, Fei Sha, William W. Cohen
- abstract@open-review(Poster): Natural language understanding tasks such as open-domain question answering often require retrieving and assimilating factual information from multiple sources. We propose to address this problem by integrating a semi-parametric representation of a large text corpus into a Transformer model as a source of factual knowledge. Specifically, our method represents knowledge with ``mention memory", a table of dense vector representations of every entity mention in a corpus. The proposed model - TOME - is a Transformer that accesses the information through internal memory layers in which each entity mention in the input passage attends to the mention memory. This approach enables synthesis of and reasoning over many disparate sources of information within a single Transformer model. In experiments using a memory of 150 million Wikipedia mentions, TOME achieves strong performance on several open-domain knowledge-intensive tasks, including the claim verification benchmarks HoVer and FEVER and several entity-based QA benchmarks. We also show that the model learns to attend to informative mentions without any direct supervision. Finally we demonstrate that the model can generalize to new unseen entities by updating the memory without retraining.

Training Data Generating Networks: Shape Reconstruction via Bi-level Optimization

- Biao Zhang, Peter Wonka
- abstract@open-review(Poster): We propose a novel 3d shape representation for 3d shape reconstruction from a single image. Rather than predicting a shape directly, we train a network to generate a training set which will be fed into another learning algorithm to define the shape. The nested optimization problem can be modeled by bi-level optimization. Specifically, the algorithms for bi-level optimization are also being used in meta learning approaches for few-shot learning. Our framework establishes a link between 3D shape analysis and few-shot learning. We combine training data generating networks with bi-level optimization algorithms to obtain a complete framework for which all components can be jointly trained. We improve upon recent work on standard benchmarks for 3d shape reconstruction.

Monotonic Differentiable Sorting Networks

- Felix Petersen, Christian Borgelt, Hilde Kuehne, Oliver Deussen
- abstract@open-review(Poster): Differentiable sorting algorithms allow training with sorting and ranking supervision, where only the ordering or ranking of samples is known. Various methods have been proposed to address this challenge, ranging from optimal transport-based differentiable Sinkhorn sorting algorithms to making classic sorting networks differentiable. One problem of current differentiable sorting methods is that they are non-monotonic. To address this issue, we propose a novel relaxation of conditional swap operations that guarantees monotonicity in differentiable sorting networks. We introduce a family of sigmoid functions and prove that they produce differentiable sorting networks that are monotonic. Monotonicity ensures that the gradients always have the correct sign, which is an advantage in gradient-based optimization. We demonstrate that monotonic differentiable sorting networks improve upon previous differentiable sorting methods.

CrowdPlay: Crowdsourcing Human Demonstrations for Offline Learning

- Matthias Gerstgrasser, Rakshit Trivedi, David C. Parkes
- abstract@open-review(Poster): Crowdsourcing has been instrumental for driving AI advances that rely on large-scale data. At the same time, reinforcement learning has seen rapid progress through benchmark environments that strike a balance between tractability and real-world complexity, such as ALE and OpenAI Gym. In this paper, we aim to fill a gap at the intersection of these two: The use of crowdsourcing to generate large-scale human demonstration data in the support of advancing research into imitation learning and offline learning. To this end, we present CrowdPlay, a complete crowdsourcing pipeline for any standard RL environment including OpenAI Gym (made available under an open-source license); a large-scale publicly available crowdsourced dataset of human gameplay demonstrations in Atari 2600 games, including multimodal behavior and human-human and human-AI multiagent data; offline learning benchmarks with extensive human data evaluation; and a detailed study of incentives, including real-time feedback to drive high quality data. We hope that this will drive the improvement in design of algorithms that account for the complexity of human, behavioral data and thereby enable a step forward in direction of effective learning for real-world settings. Our code and dataset are available at https://mgerstgrasser.github.io/crowdplay/.

Model Agnostic Interpretability for Multiple Instance Learning

- Joseph Early, Christine Evers, SArvapali Ramchurn
- abstract@open-review(Poster): In Multiple Instance Learning (MIL), models are trained using bags of instances, where only a single label is provided for each bag. A bag label is often only determined by a handful of key instances within a bag, making it difficult to interpret what information a classifier is using to make decisions. In this work, we establish the key requirements for interpreting MIL models. We then go on to develop several model-agnostic approaches that meet these requirements. Our methods are compared against existing inherently interpretable MIL models on several datasets, and achieve an increase in interpretability accuracy of up to 30%. We also examine the ability of the methods to identify interactions between instances and scale to larger datasets, improving their applicability to real-world problems.

FastSHAP: Real-Time Shapley Value Estimation

- Neil Jethani, Mukund Sudarshan, Ian Connick Covert, Su-In Lee, Rajesh Ranganath
- abstract@open-review(Poster): Although Shapley values are theoretically appealing for explaining black-box models, they are costly to calculate and thus impractical in settings that involve large, high-dimensional models. To remedy this issue, we introduce FastSHAP, a new method for estimating Shapley values in a single forward pass using a learned explainer model. To enable efficient training without requiring ground truth Shapley values, we develop an approach to train FastSHAP via stochastic gradient descent using a weighted least-squares objective function. In our experiments with tabular and image datasets, we compare FastSHAP to existing estimation approaches and find that it generates accurate explanations with an orders-of-magnitude speedup.

When, Why, and Which Pretrained GANs Are Useful?

- Timofey Grigoryev, Andrey Voynov, Artem Babenko
- abstract@open-review(Poster): The literature has proposed several methods to finetune pretrained GANs on new datasets, which typically results in higher performance compared to training from scratch, especially in the limited-data regime. However, despite the apparent empirical benefits of GAN pretraining, its inner mechanisms were not analyzed in-depth, and understanding of its role is not entirely clear. Moreover, the essential practical details, e.g., selecting a proper pretrained GAN checkpoint, currently do not have rigorous grounding and are typically determined by trial and error.

This work aims to dissect the process of GAN finetuning. First, we show that initializing the GAN training process by a pretrained checkpoint primarily affects the model's coverage rather than the fidelity of individual samples. Second, we explicitly describe how pretrained generators and discriminators contribute to the finetuning process and explain the previous evidence on the importance of pretraining both of them. Finally, as an immediate practical benefit of our analysis, we describe a simple recipe to choose an appropriate GAN checkpoint that is the most suitable for finetuning to a particular target task. Importantly, for most of the target tasks, Imagenet-pretrained GAN, despite having poor visual quality, appears to be an excellent starting point for finetuning, resembling the typical pretraining scenario of discriminative computer vision models.

A global convergence theory for deep ReLU implicit networks via over-parameterization

- Tianxiang Gao, Hailiang Liu, Jia Liu, Hridesh Rajan, Hongyang Gao
- abstract@open-review(Poster): Implicit deep learning has received increasing attention recently due to the fact that it generalizes the recursive prediction rule of many commonly used neural network architectures. Its prediction rule is provided implicitly based on the solution of an equilibrium equation. Although a line of recent empirical studies has demonstrated its superior performances, the theoretical understanding of implicit neural networks is limited. In general, the equilibrium equation may not be well-posed during the training. As a result, there is no guarantee that a vanilla (stochastic) gradient descent (SGD) training nonlinear implicit neural networks can converge. This paper fills the gap by analyzing the gradient flow of Rectified Linear Unit (ReLU) activated implicit neural networks. For an m width implicit neural network with ReLU activation and n training samples, we show that a randomly initialized gradient descent converges to a global minimum at a linear rate for the square loss function if the implicit neural network is over-parameterized. It is worth noting that, unlike existing works on the convergence of (S)GD on finite-layer over-parameterized neural networks, our convergence results hold for implicit neural networks, where the number of layers is infinite.

Learnability Lock: Authorized Learnability Control Through Adversarial Invertible Transformations

- Weiqi Peng, Jinghui Chen
- abstract@open-review(Poster): Owing much to the revolution of information technology, recent progress of deep learning benefits incredibly from the vastly enhanced access to data available in various digital formats. Yet those publicly accessible information also raises a fundamental issue concerning Intellectual Property, that is, how to precisely control legal or illegal exploitation of a dataset for training commercial models. To tackle this issue, this paper introduces and investigates a new concept called "learnability lock" for securing the process of data authorization. In particular, we propose adversarial invertible transformation, that can be viewed as a mapping from image to image, to encrypt data samples so that they become "unlearnable" by machine learning models with negligible loss of visual features. Meanwhile, authorized clients can use a specific key to unlock the learnability of the protected dataset and train models normally. The proposed learnability lock leverages class-wise perturbation that applies a universal transformation function on data samples of the same label. This ensures that the learnability can be easily restored with a simple inverse transformation while remaining difficult to be detected or reverse-engineered. We empirically demonstrate the success and practicability of our method on visual classification tasks.

Federated Learning from Only Unlabeled Data with Class-conditional-sharing Clients

- Nan Lu, Zhao Wang, Xiaoxiao Li, Gang Niu, Qi Dou, Masashi Sugiyama
- abstract@open-review(Poster): Supervised federated learning (FL) enables multiple clients to share the trained model without sharing their labeled data. However, potential clients might even be reluctant to label their own data, which could limit the applicability of FL in practice. In this paper, we show the possibility of unsupervised FL whose model is still a classifier for predicting class labels, if the class-prior probabilities are shifted while the class-conditional distributions are shared among the unlabeled data owned by the clients. We propose federation of unsupervised learning (FedUL), where the unlabeled data are transformed into surrogate labeled data for each of the clients, a modified model is trained by supervised FL, and the wanted model is

recovered from the modified model. FedUL is a very general solution to unsupervised FL: it is compatible with many supervised FL methods, and the recovery of the wanted model can be theoretically guaranteed as if the data have been labeled. Experiments on benchmark and real-world datasets demonstrate the effectiveness of FedUL. Code is available at https://github.com/lunanbit/FedUL.

Transformer Embeddings of Irregularly Spaced Events and Their Participants

- Hongyuan Mei, Chenghao Yang, Jason Eisner
- abstract@open-review(Poster): The neural Hawkes process (Mei & Eisner, 2017) is a generative model of irregularly spaced sequences of discrete events. To handle complex domains with many event types, Mei et al. (2020a) further consider a setting in which each event in the sequence updates a deductive database of facts (via domain-specific pattern-matching rules); future events are then conditioned on the database contents. They show how to convert such a symbolic system into a neuro-symbolic continuous-time generative model, in which each database fact and possible event has a time-varying embedding that is derived from its symbolic provenance.

In this paper, we modify both models, replacing their recurrent LSTM-based architectures with flatter attention-based architectures (Vaswani et al., 2017), which are simpler and more parallelizable. This does not appear to hurt our accuracy, which is comparable to or better than that of the original models as well as (where applicable) previous attention-based methods (Zuo et al., 2020; Zhang et al., 2020a).

Fast Model Editing at Scale

- Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, Christopher D Manning
- abstract@open-review(Poster): While large pre-trained models have enabled impressive results on a variety of downstream tasks, the largest existing models still make errors, and even accurate predictions may become outdated over time. Because detecting all such failures at training time is impossible, enabling both developers and end users of such models to correct inaccurate outputs while leaving the model otherwise intact is desirable. However, the distributed, black-box nature of the representations learned by large neural networks makes producing such targeted edits difficult. If presented with only a single problematic input and new desired output, fine-tuning approaches tend to overfit; other editing algorithms are either computationally infeasible or simply ineffective when applied to very large models. To enable easy post-hoc editing at scale, we propose Model Editor Networks using Gradient Decomposition (MEND), a collection of small auxiliary editing networks that use a single desired input-output pair to make fast, local edits to a pre-trained model's behavior. MEND learns to transform the gradient obtained by standard fine-tuning, using a low-rank decomposition of the gradient to make the parameterization of this transformation tractable. MEND can be trained on a single GPU in less than a day even for 10 billion+ parameter models; once trained MEND enables rapid application of new edits to the pre-trained model. Our experiments with T5, GPT, BERT, and BART models show that MEND is the only approach to model editing that effectively edits the behavior of models with more than 10 billion parameters. Code available at https://sites.google.com/view/mend-editing.

Eigencurve: Optimal Learning Rate Schedule for SGD on Quadratic Objectives with Skewed Hessian Spectrums

- Rui Pan, Haishan Ye, Tong Zhang
- abstract@open-review(Poster): Learning rate schedulers have been widely adopted in training deep neural networks. Despite their practical importance, there is a discrepancy between its practice and its theoretical analysis. For instance, it is not known what schedules of SGD achieve best convergence, even for simple problems such as optimizing quadratic objectives. In this paper, we propose Eigencurve, the first family of learning rate schedules that can achieve minimax optimal convergence rates (up to a constant) for SGD on quadratic objectives when the eigenvalue distribution of the underlying Hessian matrix is skewed. The condition is quite common in practice. Experimental results show that Eigencurve can significantly outperform step decay in image classification tasks on CIFAR-10, especially when the number of epochs is small. Moreover, the theory inspires two simple learning rate schedulers for practical applications that can approximate eigencurve. For some problems, the optimal shape of the proposed schedulers resembles that of cosine decay, which sheds light to the success of cosine decay for such situations. For other situations, the proposed schedulers are superior to cosine decay.

An Autoregressive Flow Model for 3D Molecular Geometry Generation from Scratch

- Youzhi Luo, Shuiwang Ji
- abstract@open-review(Poster): We consider the problem of generating 3D molecular geometries from scratch. While multiple methods have been developed for generating molecular graphs, generating 3D molecular geometries from scratch is largely under-explored. In this work, we propose G-SphereNet, a novel autoregressive flow model for generating 3D molecular geometries. G-SphereNet employs a flexible sequential generation scheme by placing atoms in 3D space step-by-step. Instead of generating 3D coordinates directly, we propose to determine 3D positions of atoms by generating distances, angles and torsion angles, thereby ensuring both invariance and equivariance properties. In addition, we propose to use spherical message passing and attention mechanism for conditional information extraction. Experimental results show that G-SphereNet outperforms previous methods on random molecular geometry generation and targeted molecule discovery tasks. Our code is publicly available as part of the DIG package (https://github.com/divelab/DIG).

On Incorporating Inductive Biases into VAEs

- Ning Miao, Emile Mathieu, Siddharth N, Yee Whye Teh, Tom Rainforth
- abstract@open-review(Poster): We explain why directly changing the prior can be a surprisingly ineffective mechanism for incorporating inductive biases into variational auto-encoders (VAEs), and introduce a simple and effective alternative approach: Intermediary Latent Space VAEs (InteL-VAEs). InteL-VAEs use an intermediary set of latent variables to control the stochasticity of the encoding process, before mapping these in turn to the latent representation using a parametric function that encapsulates our desired inductive bias(es). This allows us to impose properties like sparsity or clustering on learned representations, and incorporate human knowledge into the generative model. Whereas changing the prior only indirectly encourages behavior through regularizing the encoder, InteL-VAEs are able to directly enforce desired characteristics. Moreover, they bypass the computation and encoder design issues caused by non-Gaussian priors, while allowing for additional flexibility through training of the parametric mapping function. We show that these advantages, in turn, lead to both better generative models and better representations being learned.

DiffSkill: Skill Abstraction from Differentiable Physics for Deformable Object Manipulations with Tools

- Xingyu Lin, Zhiao Huang, Yunzhu Li, Joshua B. Tenenbaum, David Held, Chuang Gan
- abstract@open-review(Poster): We consider the problem of sequential robotic manipulation of deformable objects using tools. Previous works have shown that differentiable physics simulators provide gradients to the environment state and help trajectory optimization to converge orders of magnitude faster than model-free reinforcement learning algorithms for deformable object manipulation. However, such gradient-based trajectory optimization typically requires access to the full simulator states and can only solve short-horizon, single-skill tasks due to local optima. In this work, we propose a novel framework, named DiffSkill, that uses a differentiable physics simulator for skill abstraction to solve long-horizon deformable object manipulation tasks from sensory observations. In particular, we first obtain short-horizon skills using individual tools from a gradient-based optimizer, using the full state information in a differentiable simulator; we then learn a neural skill abstractor from the demonstration trajectories which takes RGBD images as input. Finally, we plan over the skills by finding the intermediate goals and then solve long-horizon tasks. We show the advantages of our method in a new set of sequential deformable object manipulation tasks compared to previous reinforcement learning algorithms and compared to the trajectory optimizer.

On the Existence of Universal Lottery Tickets

- Rebekka Burkholz, Nilanjana Laha, Rajarshi Mukherjee, Alkis Gotovos
- abstract@open-review(Poster): The lottery ticket hypothesis conjectures the existence of sparse subnetworks of large randomly initialized deep neural networks that can be successfully trained in isolation. Recent work has experimentally observed that some of these tickets can be practically reused across a variety of tasks, hinting at some form of universality. We formalize this concept and theoretically prove that not only do such universal tickets exist but they also do not require further training. Our proofs introduce a couple of technical innovations related to pruning for strong lottery tickets, including extensions of subset sum results and a strategy to leverage higher amounts of depth. Our explicit sparse constructions of universal function families might be of independent interest, as they highlight representational benefits induced by univariate convolutional architectures.

Pre-training Molecular Graph Representation with 3D Geometry

- Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, Jian Tang
- abstract@open-review(Poster): Molecular graph representation learning is a fundamental problem in modern drug and material discovery. Molecular graphs are typically modeled by their 2D topological structures, but it has been recently discovered that 3D geometric information plays a more vital role in predicting molecular functionalities. However, the lack of 3D information in real-world scenarios has significantly impeded the learning of geometric graph representation. To cope with this challenge, we propose the Graph Multi-View Pre-training (GraphMVP) framework where self-supervised learning (SSL) is performed by leveraging the correspondence and consistency between 2D topological structures and 3D geometric views. GraphMVP effectively learns a 2D molecular graph encoder that is enhanced by richer and more discriminative 3D geometry. We further provide theoretical insights to justify the effectiveness of GraphMVP. Finally, comprehensive experiments show that GraphMVP can consistently outperform existing graph SSL methods. Code is available on GitHub: https://github.com/chao1224/GraphMVP.

PER-ETD: A Polynomially Efficient Emphatic Temporal Difference Learning Method

- Ziwei Guan, Tengyu Xu, Yingbin Liang
- abstract@open-review(Poster): Emphatic temporal difference (ETD) learning (Sutton et al., 2016) is a successful method to conduct the off-policy value function evaluation with function approximation. Although ETD has been shown to converge asymptotically to a desirable value function, it is well-known that ETD often encounters a large variance so that its sample complexity can increase exponentially fast with the number of iterations. In this work, we propose a new ETD method, called PER-ETD (i.e., PEriodically Restarted-ETD), which restarts and updates the follow-on trace only for a finite period for each iteration of the evaluation parameter. Further, PER-ETD features a design of the logarithmical increase of the restart period with the number of iterations, which guarantees the best trade-off between the variance and bias and keeps both vanishing sublinearly. We show that PER-ETD converges to the same desirable fixed point as ETD, but improves the exponential sample complexity of ETD to be polynomials. Our experiments validate the superior performance of PER-ETD and its advantage over ETD.

Taming Sparsely Activated Transformer with Stochastic Experts

- Simiao Zuo, Xiaodong Liu, Jian Jiao, Young Jin Kim, Hany Hassan, Ruofei Zhang, Jianfeng Gao, Tuo Zhao
- abstract@open-review(Poster): Sparsely activated models (SAMs), such as Mixture-of-Experts (MoE), can easily scale to have outrageously large amounts of parameters without significant increase in computational cost. However, SAMs are reported to be parameter inefficient such that larger models do not always lead to better performance. While most on-going research focuses on improving SAMs models by exploring methods of routing inputs to experts, our analysis reveals that such research might not lead to the solution we expect, i.e., the commonly-used routing methods based on gating mechanisms do not work better than randomly routing inputs to experts. In this paper, we propose a new expert-based model, THOR ($\underline{\textbf{T}}$ransformer wit$\underline{\textbf{H}}$ St$\underline{\textbf{O}}$chastic Expe$\underline{\textbf{R}}$ts). Unlike classic expert-based models, such as the Switch Transformer, experts in THOR are randomly activated for each input during training and inference. THOR models are trained using a consistency regularized loss, where experts learn not only from training data but also from other experts as teachers, such that all the experts make consistent predictions. We validate the effectiveness of THOR on machine translation tasks. Results show that THOR models are more parameter efficient in that they significantly outperform the Transformer and MoE models across various settings. For example, in multilingual translation, THOR outperforms the Switch Transformer by 2 BLEU scores, and obtains the same BLEU score as that of a state-of-the-art MoE model that is 18 times larger. Our code is publicly available at: https://github.com/microsoft/Stochastic-Mixture-of-Experts.

Hierarchical Variational Memory for Few-shot Learning Across Domains

- Yingjun Du, Xiantong Zhen, Ling Shao, Cees G. M. Snoek
- abstract@open-review(Poster): Neural memory enables fast adaptation to new tasks with just a few training samples. Existing memory models store features only from the single last layer, which does not generalize well in presence of a domain shift between training and test distributions. Rather than relying on a flat memory, we propose a hierarchical alternative that stores features at different semantic levels. We introduce a hierarchical prototype model, where each level of the prototype fetches corresponding information from the hierarchical memory. The model is endowed with the ability to flexibly rely on features at different semantic levels if the domain shift circumstances so demand. We meta-learn the model by a newly derived hierarchical variational inference framework, where hierarchical memory and prototypes are jointly optimized. To explore and exploit the importance of different semantic levels, we further propose to learn the weights associated with the prototype at each level in a data-driven way, which enables the model to adaptively choose the most generalizable features. We conduct thorough ablation studies to demonstrate the effectiveness of each component in our model. The new state-of-the-art performance on cross-domain and competitive performance on traditional few-shot classification further substantiates the benefit of hierarchical variational memory.

Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction

- Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, Abdelrahman Mohamed
- abstract@open-review(Poster): Video recordings of speech contain correlated audio and visual information, providing a strong signal for speech representation learning from the speakerâ€™s lip movements and the produced sound. We introduce Audio-Visual Hidden Unit BERT (AV-HuBERT), a self-supervised representation learning framework for audio-visual speech, which masks multi-stream video input and predicts automatically discovered and iteratively refined multimodal hidden units. AV-HuBERT learns powerful audio-visual speech representation benefiting both lip-reading and automatic speech recognition. On the largest public lip-reading benchmark LRS3 (433 hours), AV-HuBERT achieves 32.5% WER with only 30 hours of labeled data, outperforming the former state-of-the-art approach (33.6%) trained with a thousand times more transcribed video data (31K hours) (Makino et al., 2019). The lip-reading WER is further reduced to 26.9% when using all 433 hours of labeled data from LRS3 and combined with self-training. Using our audio-visual representation on the same benchmark for audio-only speech recognition leads to a 40% relative WER reduction over the state-of-the-art performance (1.3% vs 2.3%). Our code and models are available at https://github.com/facebookresearch/av_hubert.

An Explanation of In-context Learning as Implicit Bayesian Inference

- Sang Michael Xie, Aditi Raghunathan, Percy Liang, Tengyu Ma

- abstract@open-review(Poster): Large language models (LMs) such as GPT-3 have the surprising ability to do in-context learning, where the model learns to do a downstream task simply by conditioning on a prompt consisting of input-output examples. The LM learns from these examples without being explicitly pretrained to learn. Thus, it is unclear what enables in-context learning. In this paper, we study how in-context learning can emerge when pretraining documents have long-range coherence. Here, the LM must infer a latent document-level concept to generate coherent next tokens during pretraining. At test time, in-context learning occurs when the LM also infers a shared latent concept between examples in a prompt. We prove when this occurs despite a distribution mismatch between prompts and pretraining data in a setting where the pretraining distribution is a mixture of HMMs. In contrast to messy large-scale datasets used to train LMs capable of in-context learning, we generate a small-scale synthetic dataset (GINC) where Transformers and LSTMs both exhibit in-context learning. Beyond the theory, experiments on GINC exhibit large-scale real-world phenomena including improved in-context performance with model scaling (despite the same pretraining loss), sensitivity to example order, and instances where zero-shot is better than few-shot in-context learning.

Differentiable Scaffolding Tree for Molecule Optimization

- Tianfan Fu, Wenhao Gao, Cao Xiao, Jacob Yasonik, Connor W. Coley, Jimeng Sun
- abstract@open-review(Poster): The structural design of functional molecules, also called molecular optimization, is an essential chemical science and engineering task with important applications, such as drug discovery. Deep generative models and combinatorial optimization methods achieve initial success but still struggle with directly modeling discrete chemical structures and often heavily rely on brute-force enumeration. The challenge comes from the discrete and non-differentiable nature of molecule structures. To address this, we propose differentiable scaffolding tree (DST) that utilizes a learned knowledge network to convert discrete chemical structures to locally differentiable ones. DST enables a gradient-based optimization on a chemical graph structure by back-propagating the derivatives from the target properties through a graph neural network (GNN). Our empirical studies show the gradient-based molecular optimizations are both effective and sample efficient (in terms of oracle calling number). Furthermore, the learned graph parameters can also provide an explanation that helps domain experts understand the model output. The code repository (including processed data, trained model, demonstration, molecules with the highest property) is available at https://github.com/futianfan/DST.

Eliminating Sharp Minima from SGD with Truncated Heavy-tailed Noise

- Xingyu Wang, Sewoong Oh, Chang-Han Rhee
- abstract@open-review(Poster): The empirical success of deep learning is often attributed to SGDâ€™s mysterious ability to avoid sharp local minima in the loss landscape, as sharp minima are known to lead to poor generalization. Recently, empirical evidence of heavy-tailed gradient noise was reported in many deep learning tasks; and it was shown in (Simsekli et al., 2019a;b) that SGD can escape sharp local minima under the presence of such heavy-tailed gradient noise, providing a partial solution to the mystery. In this work, we analyze a popular variant of SGD where gradients are truncated above a fixed threshold. We show that it achieves a stronger notion of avoiding sharp minima: it can effectively eliminate sharp local minima entirely from its training trajectory. We characterize the dynamics of truncated SGD driven by heavy-tailed noises. First, we show that the truncation threshold and width of the attraction field dictate the order of the first exit time from the associated local minimum. Moreover, when the objective function satisfies appropriate structural conditions, we prove that as the learning rate decreases, the dynamics of the heavy-tailed truncated SGD closely resemble those of a continuous-time Markov chain that never visits any sharp minima. Real data experiments on deep learning confirm our theoretical prediction that heavy-tailed SGD with gradient clipping finds a flatter local minima and achieves better generalization.

Learning Fast, Learning Slow: A General Continual Learning Method based on Complementary Learning System

- Elahe Arani, Fahad Sarfraz, Bahram Zonooz
- abstract@open-review(Poster): Humans excel at continually learning from an ever-changing environment whereas it remains a challenge for deep neural networks which exhibit catastrophic forgetting. The complementary learning system (CLS) theory suggests that the interplay between rapid instance-based learning and slow structured learning in the brain is crucial for accumulating and retaining knowledge. Here, we propose CLS-ER, a novel dual memory experience replay (ER) method which maintains short-term and long-term semantic memories that interact with the episodic memory. Our method employs an effective replay mechanism whereby new knowledge is acquired while aligning the decision boundaries with the semantic memories. CLS-ER does not utilize the task boundaries or make any assumption about the distribution of the data which makes it versatile and suited for ``general continual learning''. Our approach achieves state-of-the-art performance on standard benchmarks as well as more realistic general continual learning settings.

FedChain: Chained Algorithms for Near-optimal Communication Cost in Federated Learning

- Charlie Hou, Kiran Koshy Thekumparampil, Giulia Fanti, Sewoong Oh
- abstract@open-review(Poster): Federated learning (FL) aims to minimize the communication complexity of training a model over heterogeneous data distributed across many clients. A common approach is local methods, where clients take multiple optimization steps over local data before communicating with the server (e.g., FedAvg). Local methods can exploit similarity between clients' data. However, in existing analyses, this comes at the cost of slow convergence in terms of the dependence on the number of communication rounds R. On the other hand, global methods, where clients simply return a gradient vector in each round (e.g., SGD), converge faster in terms of R but fail to exploit the similarity between clients even when clients are homogeneous. We propose FedChain, an algorithmic framework that combines the strengths of local methods and global methods to achieve fast convergence in terms of R while leveraging the similarity between clients. Using FedChain, we instantiate algorithms that improve upon previously known rates in the general convex and PL settings, and are near-optimal (via an algorithm-independent lower bound that we show) for problems that satisfy strong convexity. Empirical results support this theoretical gain over existing methods.

What Do We Mean by Generalization in Federated Learning?

- Honglin Yuan, Warren Richard Morningstar, Lin Ning, Karan Singhal
- abstract@open-review(Poster): Federated learning data is drawn from a distribution of distributions: clients are drawn from a meta-distribution, and their data are drawn from local data distributions. Generalization studies in federated learning should separate performance gaps from unseen client data (out-of-sample gap) from performance gaps from unseen client distributions (participation gap). In this work, we propose a framework for disentangling these performance gaps. Using this framework, we observe and explain differences in behavior across natural and synthetic federated datasets, indicating that dataset synthesis strategy can be important for realistic simulations of generalization in federated learning. We propose a semantic synthesis strategy that enables realistic simulation without naturally partitioned data. Informed by our ï¬ ndings, we call out community suggestions for future federated learning works.

Frequency-aware SGD for Efficient Embedding Learning with Provable Benefits

- Yan Li, Dhruv Choudhary, Xiaohan Wei, Baichuan Yuan, Bhargav Bhushanam, Tuo Zhao, Guanghui Lan
- abstract@open-review(Poster): Embedding learning has found widespread applications in recommendation systems and natural language modeling, among other domains. To learn quality embeddings efficiently, adaptive learning rate algorithms have demonstrated superior empirical performance over SGD, largely accredited to their token-dependent learning rate. However, the underlying mechanism for the efficiency of token-dependent learning rate remains underexplored. We show that incorporating frequency information of tokens in the embedding learning problems leads to provably efficient algorithms, and demonstrate that common adaptive algorithms implicitly exploit the frequency information to a large extent. Specifically, we propose (Counter-based) Frequency-aware Stochastic Gradient Descent, which applies a frequency-dependent learning rate for each token, and exhibits provable

speed-up compared to SGD when the token distribution is imbalanced. Empirically, we show the proposed algorithms are able to improve or match the performance of adaptive algorithms on benchmark recommendation tasks and a large-scale industrial recommendation system, closing the performance gap between SGD and adaptive algorithms. Our results are the first to show token-dependent learning rate provably improves convergence for non-convex embedding learning problems.

Learning Curves for Gaussian Process Regression with Power-Law Priors and Targets

- Hui Jin, Pradeep Kr. Banerjee, Guido Montufar
- abstract@open-review(Poster): We characterize the power-law asymptotics of learning curves for Gaussian process regression (GPR) under the assumption that the eigenspectrum of the prior and the eigenexpansion coefficients of the target function follow a power law. Under similar assumptions, we leverage the equivalence between GPR and kernel ridge regression (KRR) to show the generalization error of KRR. Infinitely wide neural networks can be related to GPR with respect to the neural network GP kernel and the neural tangent kernel, which in several cases is known to have a power-law spectrum. Hence our methods can be applied to study the generalization error of infinitely wide neural networks. We present toy experiments demonstrating the theory.

Fast topological clustering with Wasserstein distance

- Tananun Songdechakraiwut, Bryan M Krause, Matthew I Banks, Kirill V Nourski, Barry D Van Veen
- abstract@open-review(Poster): The topological patterns exhibited by many real-world networks motivate the development of topology-based methods for assessing the similarity of networks. However, extracting topological structure is difficult, especially for large and dense networks whose node degrees range over multiple orders of magnitude. In this paper, we propose a novel and computationally practical topological clustering method that clusters complex networks with intricate topology using principled theory from persistent homology and optimal transport. Such networks are aggregated into clusters through a centroid-based clustering strategy based on both their topological and geometric structure, preserving correspondence between nodes in different networks. The notions of topological proximity and centroid are characterized using a novel and efficient approach to computation of the Wasserstein distance and barycenter for persistence barcodes associated with connected components and cycles. The proposed method is demonstrated to be effective using both simulated networks and measured functional brain networks.

Autonomous Reinforcement Learning: Formalism and Benchmarking

- Archit Sharma, Kelvin Xu, Nikhil Sardana, Abhishek Gupta, Karol Hausman, Sergey Levine, Chelsea Finn
- abstract@open-review(Poster): Reinforcement learning (RL) provides a naturalistic framing for learning through trial and error, which is appealing both because of its simplicity and effectiveness and because of its resemblance to how humans and animals acquire skills through experience. However, real-world embodied learning, such as that performed by humans and animals, is situated in a continual, non-episodic world, whereas common benchmark tasks in RL are episodic, with the environment resetting between trials to provide the agent with multiple attempts. This discrepancy presents a major challenge when we attempt to take RL algorithms developed for episodic simulated environments and run them on real-world platforms, such as robots. In this paper, we aim to address this discrepancy by laying out a framework for Autonomous Reinforcement Learning (ARL): reinforcement learning where the agent not only learns through its own experience, but also contends with lack of human supervision to reset between trials. We introduce a simulated benchmark EARL based on this framework, containing a set of diverse and challenging simulated tasks reflective of the hurdles introduced to learning when only a minimal reliance on extrinsic intervention can be assumed. We show that standard approaches to episodic RL and existing approaches struggle as interventions are minimized, underscoring the need for developing new algorithms for reinforcement learning with a greater focus on autonomy.

GRAND++: Graph Neural Diffusion with A Source Term

- Matthew Thorpe, Tan Minh Nguyen, Hedi Xia, Thomas Strohmer, Andrea Bertozzi, Stanley Osher, Bao Wang
- abstract@open-review(Poster): We propose GRAph Neural Diffusion with a source term (GRAND++) for graph deep learning with a limited number of labeled nodes, i.e., low-labeling rate. GRAND++ is a class of continuous-depth graph deep learning architectures whose theoretical underpinning is the diffusion process on graphs with a source term. The source term guarantees two interesting theoretical properties of GRAND++: (i) the representation of graph nodes, under the dynamics of GRAND++, will not converge to a constant vector over all nodes even as the time goes to infinity, which mitigates the over-smoothing issue of graph neural networks and enables graph learning in very deep architectures. (ii) GRAND++ can provide accurate classification even when the model is trained with a very limited number of labeled training data. We experimentally verify the above two advantages on various graph deep learning benchmark tasks, showing a significant improvement over many existing graph neural networks.

Case-based reasoning for better generalization in textual reinforcement learning

- Mattia Atzeni, Shehzaad Zuzar Dhuliawala, Keerthiram Murugesan, Mrinmaya Sachan
- abstract@open-review(Poster): Text-based games (TBG) have emerged as promising environments for driving research in grounded language understanding and studying problems like generalization and sample efficiency. Several deep reinforcement learning (RL) methods with varying architectures and learning schemes have been proposed for TBGs. However, these methods fail to generalize efficiently, especially under distributional shifts. In a departure from deep RL approaches, in this paper, we propose a general method inspired by case-based reasoning to train agents and generalize out of the training distribution. The case-based reasoner collects instances of positive experiences from the agent's interaction with the world and later reuses the collected experiences to act efficiently. The method can be used in conjunction with any existing on-policy neural agent introduced in the literature for TBGs. Our experiments show that the proposed approach consistently improves existing methods, obtains good out-of-distribution generalization and achieves new state-of-the-art results on widely used environments.

Neural Deep Equilibrium Solvers

- Shaojie Bai, Vladlen Koltun, J Zico Kolter
- abstract@open-review(Poster): A deep equilibrium (DEQ) model abandons traditional depth by solving for the fixed point of a single nonlinear layer f_θ. This structure enables decoupling the internal structure of the layer (which controls representational capacity) from how the fixed point is actually computed (which impacts inference-time efficiency), which is usually via classic techniques such as Broyden's method or Anderson acceleration. In this paper, we show that one can exploit such decoupling and substantially enhance this fixed point computation using a custom neural solver. Specifically, our solver uses a parameterized network to both guess an initial value of the optimization and perform iterative updates, in a method that generalizes a learnable form of Anderson acceleration and can be trained end-to-end in an unsupervised manner. Such a solution is particularly well suited to the implicit model setting, because inference in these models requires repeatedly solving for a fixed point of the same nonlinear layer for different inputs, a task at which our network excels. Our experiments show that these neural equilibrium solvers are fast to train (only taking an extra 0.9-1.1% over the original DEQ's training time), require few additional parameters (1-3% of the original model size), yet lead to a $2\times$ speedup in DEQ network inference without any degradation in accuracy across numerous domains and tasks.

A Theoretical Analysis on Feature Learning in Neural Networks: Emergence from Inputs and Advantage over Fixed Features

- Zhenmei Shi, Junyi Wei, Yingyu Liang
- abstract@open-review(Poster): An important characteristic of neural networks is their ability to learn representations of the input data with effective features for prediction, which is believed to be a key factor to their superior empirical performance. To better understand the source and benefit of feature learning in neural networks, we consider learning problems motivated by practical data, where the labels are determined by a set of class relevant patterns and the inputs are generated from these along with some background patterns. We prove that neural networks trained by gradient descent can succeed on these problems. The success relies on the emergence and improvement of effective features, which are learned among exponentially many candidates efficiently by exploiting the data (in particular, the structure of the input distribution). In contrast, no linear models on data-independent features of polynomial sizes can learn to as good errors. Furthermore, if the specific input structure is removed, then no polynomial algorithm in the Statistical Query model can learn even weakly. These results provide theoretical evidence showing that feature learning in neural networks depends strongly on the input structure and leads to the superior performance. Our preliminary experimental results on synthetic and real data also provide positive support.

CADDA: Class-wise Automatic Differentiable Data Augmentation for EEG Signals

- CÃ©dric Rommel, Thomas Moreau, Joseph Paillard, Alexandre Gramfort
- abstract@open-review(Poster): Data augmentation is a key element of deep learning pipelines, as it informs the network during training about transformations of the input data that keep the label unchanged. Manually finding adequate augmentation methods and parameters for a given pipeline is however rapidly cumbersome. In particular, while intuition can guide this decision for images, the design and choice of augmentation policies remains unclear for more complex types of data, such as neuroscience signals. Besides, class-dependent augmentation strategies have been surprisingly unexplored in the literature, although it is quite intuitive: changing the color of a car image does not change the object class to be predicted, but doing the same to the picture of an orange does. This paper investigates gradient-based automatic data augmentation algorithms amenable to class-wise policies with exponentially larger search spaces. Motivated by supervised learning applications using EEG signals for which good augmentation policies are mostly unknown, we propose a new differentiable relaxation of the problem. In the class-agnostic setting, results show that our new relaxation leads to optimal performance with faster training than competing gradient-based methods, while also outperforming gradient-free methods in the class-wise setting. This work proposes also novel differentiable augmentation operations relevant for sleep stage classification.

Label Leakage and Protection in Two-party Split Learning

- Oscar Li, Jiankai Sun, Xin Yang, Weihao Gao, Hongyi Zhang, Junyuan Xie, Virginia Smith, Chong Wang
- abstract@open-review(Poster): Two-party split learning is a popular technique for learning a model across feature-partitioned data. In this work, we explore whether it is possible for one party to steal the private label information from the other party during split training, and whether there are methods that can protect against such attacks. Specifically, we first formulate a realistic threat model and propose a privacy loss metric to quantify label leakage in split learning. We then show that there exist two simple yet effective methods within the threat model that can allow one party to accurately recover private ground-truth labels owned by the other party. To combat these attacks, we propose several random perturbation techniques, including $\texttt{Marvell}$, an approach that strategically finds the structure of the noise perturbation by minimizing the amount of label leakage (measured through our quantification metric) of a worst-case adversary. We empirically demonstrate the effectiveness of our protection techniques against the identified attacks, and show that $\texttt{Marvell}$ in particular has improved privacy-utility tradeoffs relative to baseline approaches.

Semi-relaxed Gromov-Wasserstein divergence and applications on graphs

- CÃ©dric Vincent-Cuaz, RÃ©mi Flamary, Marco Corneli, Titouan Vayer, Nicolas Courty
- abstract@open-review(Poster): Comparing structured objects such as graphs is a fundamental operation involved in many learning tasks. To this end, the Gromov-Wasserstein (GW) distance, based on Optimal Transport (OT), has proven to be successful in handling the specific nature of the associated objects. More specifically, through the nodes connectivity relations, GW operates on graphs, seen as probability measures over specific spaces. At the core of OT is the idea of conservation of mass, which imposes a coupling between all the nodes from the two considered graphs. We argue in this paper that this property can be detrimental for tasks such as graph dictionary or partition learning, and we relax it by proposing a new semi-relaxed Gromov-Wasserstein divergence. Aside from immediate computational benefits, we discuss its properties, and show that it can lead to an efficient graph dictionary learning algorithm. We empirically demonstrate its relevance for complex tasks on graphs such as partitioning, clustering and completion.

CodeTrek: Flexible Modeling of Code using an Extensible Relational Representation

- Pardis Pashakhanloo, Aaditya Naik, Yuepeng Wang, Hanjun Dai, Petros Maniatis, Mayur Naik
- abstract@open-review(Poster): Designing a suitable representation for code-reasoning tasks is challenging in aspects such as the kinds of program information to model, how to combine them, and how much context to consider. We propose CodeTrek, a deep learning approach that addresses these challenges by representing codebases as databases that conform to rich relational schemas. The relational representation not only allows CodeTrek to uniformly represent diverse kinds of program information, but also to leverage program-analysis queries to derive new semantic relations, which can be readily incorporated without further architectural engineering. CodeTrek embeds this relational representation using a set of walks that can traverse different relations in an unconstrained fashion, and incorporates all relevant attributes along the way. We evaluate CodeTrek on four diverse and challenging Python tasks: variable misuse, exception prediction, unused definition, and variable shadowing. CodeTrek achieves an accuracy of 91%, 63%, 98%, and 94% on these tasks respectively, and outperforms state-of-the-art neural models by 2-19% points.

Bridging Recommendation and Marketing via Recurrent Intensity Modeling

- Yifei Ma, Ge Liu, Anoop Deoras
- abstract@open-review(Poster): This paper studies some under-explored connections between personalized recommendation and marketing systems. Obviously, these two systems are different, in two main ways. Firstly, personalized item-recommendation (ItemRec) is user-centric, whereas marketing recommends the best user-state segments (UserRec) on behalf of its item providers. (We treat different temporal states of the same user as separate marketing opportunities.) To overcome this difference, we realize a novel connection to Marked-Temporal Point Processes (MTPPs), where we view both problems as different projections from a unified temporal intensity model for all user-item pairs. Correspondingly, we derive Recurrent Intensity Models (RIMs) to extend from recurrent ItemRec models with minimal changes. The second difference between recommendation and marketing is in the temporal domains where they operate. While recommendation demands immediate responses in real-time, marketing campaigns are often long-term, setting goals to cover a given percentage of all opportunities for a given item in a given period of time. We formulate both considerations into a constrained optimization problem we call online match (OnlnMtch) and derive a solution we call Dual algorithm. Simply put, Dual modifies the real-time ItemRec scores such that the marketing constraints can be met with least compromises in user-centric utilities. Finally, our connections between recommendation and marketing may lead to novel applications. We run experiments where we use marketing as an alternative to cold-start item exploration, by setting a minimal-exposure constraint for every item in the audience base. Our experiments are available at \url{https://github.com/awslabs/recurrent-intensity-model-experiments}

Sparse Attention with Learning to Hash

- Zhiqing Sun, Yiming Yang, Shinjae Yoo
- abstract@open-review(Poster): Transformer has become ubiquitous in sequence modeling tasks. As a key component of Transformer, self-attention does not scale to long sequences due to its quadratic time and space complexity with respect to the sequence length. To tackle this problem, recent work

developed dynamic attention sparsification techniques based on Approximate Nearest Neighbor (ANN) methods, where similar queries and keys are allocated to the same hash bucket with high probability. However, the effectiveness of those ANN methods relies on the assumption that queries and keys should lie in the same space, which is not well justified. Besides, some of the ANN methods such as Locality-Sensitive Hashing (LSH) are randomized and cannot fully utilize the available real data distributions. To overcome these issues, this paper proposes a new strategy for sparse attention, namely LHA (Learning-to-Hash Attention), which directly learns separate parameterized hash functions for queries and keys, respectively. Another advantage of LHA is that it does not impose extra constraints for queries and keys, which makes it applicable to the wide range of pre-trained Transformer models. Our experiments on evaluation of the WikiText-103 dataset for language modeling, the GLUE benchmark for natural language understanding, and the Lang-Range-Arena benchmark for multiple tasks (text/image classification, retrieval, etc.) show the superior performance of LHA over other strong Transformer variants.

Controlling the Complexity and Lipschitz Constant improves Polynomial Nets

- Zhenyu Zhu, Fabian Latorre, Grigorios Chrysos, Volkan Cevher
- abstract@open-review(Poster): While the class of Polynomial Nets demonstrates comparable performance to neural networks (NN), it currently has neither theoretical generalization characterization nor robustness guarantees. To this end, we derive new complexity bounds for the set of Coupled CP-Decomposition (CCP) and Nested Coupled CP-decomposition (NCP) models of Polynomial Nets in terms of the ℓ_∞-operator-norm and the ℓ_2-operator norm. In addition, we derive bounds on the Lipschitz constant for both models to establish a theoretical certificate for their robustness. The theoretical results enable us to propose a principled regularization scheme that we also evaluate experimentally and show that it improves the accuracy as well as the robustness of the models to adversarial perturbations. We showcase how this regularization can be combined with adversarial training, resulting in further improvements.

Finding an Unsupervised Image Segmenter in each of your Deep Generative Models

- Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, Andrea Vedaldi
- abstract@open-review(Poster): Recent research has shown that numerous human-interpretable directions exist in the latent space of GANs. In this paper, we develop an automatic procedure for finding directions that lead to foreground-background image separation, and we use these directions to train an image segmentation model without human supervision. Our method is generator-agnostic, producing strong segmentation results with a wide range of different GAN architectures. Furthermore, by leveraging GANs pretrained on large datasets such as ImageNet, we are able to segment images from a range of domains without further training or finetuning. Evaluating our method on image segmentation benchmarks, we compare favorably to prior work while using neither human supervision nor access to the training data. Broadly, our results demonstrate that automatically extracting foreground-background structure from pretrained deep generative models can serve as a remarkably effective substitute for human supervision.

Solving Inverse Problems in Medical Imaging with Score-Based Generative Models

- Yang Song, Liyue Shen, Lei Xing, Stefano Ermon
- abstract@open-review(Poster): Reconstructing medical images from partial measurements is an important inverse problem in Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). Existing solutions based on machine learning typically train a model to directly map measurements to medical images, leveraging a training dataset of paired images and measurements. These measurements are typically synthesized from images using a fixed physical model of the measurement process, which hinders the generalization capability of models to unknown measurement processes. To address this issue, we propose a fully unsupervised technique for inverse problem solving, leveraging the recently introduced score-based generative models. Specifically, we first train a score-based generative model on medical images to capture their prior distribution. Given measurements and a physical model of the measurement process at test time, we introduce a sampling method to reconstruct an image consistent with both the prior and the observed measurements. Our method does not assume a fixed measurement process during training, and can thus be flexibly adapted to different measurement processes at test time. Empirically, we observe comparable or better performance to supervised learning techniques in several medical imaging tasks in CT and MRI, while demonstrating significantly better generalization to unknown measurement processes.

BDDM: Bilateral Denoising Diffusion Models for Fast and High-Quality Speech Synthesis

- Max W. Y. Lam, Jun Wang, Dan Su, Dong Yu
- abstract@open-review(Poster): Diffusion probabilistic models (DPMs) and their extensions have emerged as competitive generative models yet confront challenges of efficient sampling. We propose a new bilateral denoising diffusion model (BDDM) that parameterizes both the forward and reverse processes with a schedule network and a score network, which can train with a novel bilateral modeling objective. We show that the new surrogate objective can achieve a lower bound of the log marginal likelihood tighter than a conventional surrogate. We also find that BDDM allows inheriting pre-trained score network parameters from any DPMs and consequently enables speedy and stable learning of the schedule network and optimization of a noise schedule for sampling. Our experiments demonstrate that BDDMs can generate high-fidelity audio samples with as few as three sampling steps. Moreover, compared to other state-of-the-art diffusion-based neural vocoders, BDDMs produce comparable or higher quality samples indistinguishable from human speech, notably with only seven sampling steps (143x faster than WaveGrad and 28.6x faster than DiffWave). We release our code at https://github.com/tencent-ailab/bddm.

Sample Efficient Stochastic Policy Extragradient Algorithm for Zero-Sum Markov Game

- Ziyi Chen, Shaocong Ma, Yi Zhou
- abstract@open-review(Poster): Two-player zero-sum Markov game is a fundamental problem in reinforcement learning and game theory. Although many algorithms have been proposed for solving zero-sum Markov games in the existing literature, many of them either require a full knowledge of the environment or are not sample-efficient. In this paper, we develop a fully decentralized and sample-efficient stochastic policy extragradient algorithm for solving tabular zero-sum Markov games. In particular, our algorithm utilizes multiple stochastic estimators to accurately estimate the value functions involved in the stochastic updates, and leverages entropy regularization to accelerate the convergence. Specifically, with a proper entropy-regularization parameter, we prove that the stochastic policy extragradient algorithm has a sample complexity of the order $\widetilde{\mathcal{O}}(\frac{A_{\max}}{\mu_{\text{min}}\epsilon^{5.5}(1-\gamma)^{13.5}})$ for finding a solution that achieves ϵ-Nash equilibrium duality gap, where $A_{\max}$ is the maximum number of actions between the players, $\mu_{\min}$ is the lower bound of state stationary distribution, and γ is the discount factor. Such a sample complexity result substantially improves the state-of-the-art complexity result.

The Uncanny Similarity of Recurrence and Depth

- Avi Schwarzschild, Arjun Gupta, Amin Ghiasi, Micah Goldblum, Tom Goldstein
- abstract@open-review(Poster): It is widely believed that deep neural networks contain layer specialization, wherein networks extract hierarchical features representing edges and patterns in shallow layers and complete objects in deeper layers. Unlike common feed-forward models that have distinct filters at each layer, recurrent networks reuse the same parameters at various depths. In this work, we observe that recurrent models exhibit the same hierarchical behaviors and the same performance benefits as depth despite reusing the same filters at every recurrence. By training models of various feed-forward and recurrent architectures on several datasets for image classification as well as maze solving, we show that recurrent networks have the ability to closely emulate the behavior of non-recurrent deep models, often doing so with far fewer parameters.

Implicit Bias of Adversarial Training for Deep Neural Networks

- Bochen Lv, Zhanxing Zhu
- abstract@open-review(Poster): We provide theoretical understandings of the implicit bias imposed by adversarial training for homogeneous deep neural networks without any explicit regularization. In particular, for deep linear networks adversarially trained by gradient descent on a linearly separable dataset, we prove that the direction of the product of weight matrices converges to the direction of the max-margin solution of the original dataset. Furthermore, we generalize this result to the case of adversarial training for non-linear homogeneous deep neural networks without the linear separability of the dataset. We show that, when the neural network is adversarially trained with ℓ_2 or ℓ_{∞} FGSM, FGM and PGD perturbations, the direction of the limit point of normalized parameters of the network along the trajectory of the gradient flow converges to a KKT point of a constrained optimization problem that aims to maximize the margin for adversarial examples. Our results theoretically justify the longstanding conjecture that adversarial training modifies the decision boundary by utilizing adversarial examples to improve robustness, and potentially provides insights for designing new robust training strategies.

Mastering Visual Continuous Control: Improved Data-Augmented Reinforcement Learning

- Denis Yarats, Rob Fergus, Alessandro Lazaric, Lerrel Pinto
- abstract@open-review(Poster): We present DrQ-v2, a model-free reinforcement learning (RL) algorithm for visual continuous control. DrQ-v2 builds on DrQ, an off-policy actor-critic approach that uses data augmentation to learn directly from pixels. We introduce several improvements that yield state-of-the-art results on the DeepMind Control Suite. Notably, DrQ-v2 is able to solve complex humanoid locomotion tasks directly from pixel observations, previously unattained by model-free RL. DrQ-v2 is conceptually simple, easy to implement, and provides significantly better computational footprint compared to prior work, with the majority of tasks taking just 8 hours to train on a single GPU. Finally, we publicly release DrQ-v2 's implementation to provide RL practitioners with a strong and computationally efficient baseline.

πBO: Augmenting Acquisition Functions with User Beliefs for Bayesian Optimization

- Carl Hvarfner, Danny Stoll, Artur Souza, Marius Lindauer, Frank Hutter, Luigi Nardi
- abstract@open-review(Poster): Bayesian optimization (BO) has become an established framework and popular tool for hyperparameter optimization (HPO) of machine learning (ML) algorithms. While known for its sample-efficiency, vanilla BO can not utilize readily available prior beliefs the practitioner has on the potential location of the optimum. Thus, BO disregards a valuable source of information, reducing its appeal to ML practitioners. To address this issue, we propose πBO, an acquisition function generalization which incorporates prior beliefs about the location of the optimum in the form of a probability distribution, provided by the user. In contrast to previous approaches, πBO is conceptually simple and can easily be integrated with existing libraries and many acquisition functions. We provide regret bounds when πBO is applied to the common Expected Improvement acquisition function and prove convergence at regular rates independently of the prior. Further, our experiments show that πBO outperforms competing approaches across a wide suite of benchmarks and prior characteristics. We also demonstrate that πBO improves on the state-of-the-art performance for a popular deep learning task, with a $12.5\times$ time-to-accuracy speedup over prominent BO approaches.

A Generalized Weighted Optimization Method for Computational Learning and Inversion

- Kui Ren, Yunan Yang, BjÃ¶rn Engquist
- abstract@open-review(Poster): The generalization capacity of various machine learning models exhibits different phenomena in the under- and over-parameterized regimes. In this paper, we focus on regression models such as feature regression and kernel regression and analyze a generalized weighted least-squares optimization method for computational learning and inversion with noisy data. The highlight of the proposed framework is that we allow weighting in both the parameter space and the data space. The weighting scheme encodes both a priori knowledge on the object to be learned and a strategy to weight the contribution of different data points in the loss function. Here, we characterize the impact of the weighting scheme on the generalization error of the learning method, where we derive explicit generalization errors for the random Fourier feature model in both the under- and over-parameterized regimes. For more general feature maps, error bounds are provided based on the singular values of the feature matrix. We demonstrate that appropriate weighting from prior knowledge can improve the generalization capability of the learned model.

DriPP: Driven Point Processes to Model Stimuli Induced Patterns in M/EEG Signals

- CÃ©dric Allain, Alexandre Gramfort, Thomas Moreau
- abstract@open-review(Poster): The quantitative analysis of non-invasive electrophysiology signals from electroencephalography (EEG) and magnetoencephalography (MEG) boils down to the identification of temporal patterns such as evoked responses, transient bursts of neural oscillations but also blinks or heartbeats for data cleaning. Several works have shown that these patterns can be extracted efficiently in an unsupervised way, e.g., using Convolutional Dictionary Learning. This leads to an event-based description of the data. Given these events, a natural question is to estimate how their occurrences are modulated by certain cognitive tasks and experimental manipulations. To address it, we propose a point process approach. While point processes have been used in neuroscience in the past, in particular for single cell recordings (spike trains), techniques such as Convolutional Dictionary Learning make them amenable to human studies based on EEG/MEG signals. We develop a novel statistical point process model â€“ called driven temporal point processes (DriPP) â€“ where the intensity function of the point process model is linked to a set of point processes corresponding to stimulation events. We derive a fast and principled expectation-maximization algorithm to estimate the parameters of this model. Simulations reveal that model parameters can be identified from long enough signals. Results on standard MEG datasets demonstrate that our methodology reveals event-related neural responses â€“ both evoked and induced â€“ and isolates non-task specific temporal patterns.

Stiffness-aware neural network for learning Hamiltonian systems

- SENWEI Liang, Zhongzhan Huang, Hong Zhang
- abstract@open-review(Poster): We propose stiffness-aware neural network (SANN), a new method for learning Hamiltonian dynamical systems from data. SANN identifies and splits the training data into stiff and nonstiff portions based on a stiffness-aware index, a simple, yet effective metric we introduce to quantify the stiffness of the dynamical system. This classification along with a resampling technique allows us to apply different time integration strategies such as step size adaptation to better capture the dynamical characteristics of the Hamiltonian vector fields. We evaluate SANN on complex physical systems including a three-body problem and billiard model. We show that SANN is more stable and can better preserve energy when compared with the state-of-the-art methods, leading to significant improvement in accuracy.

CoST: Contrastive Learning of Disentangled Seasonal-Trend Representations for Time Series Forecasting

- Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, Steven Hoi
- abstract@open-review(Poster): Deep learning has been actively studied for time series forecasting, and the mainstream paradigm is based on the end-to-end training of neural network architectures, ranging from classical LSTM/RNNs to more recent TCNs and Transformers. Motivated by the recent success of representation learning in computer vision and natural language processing, we argue that a more promising paradigm for time series forecasting, is to first learn disentangled feature representations, followed by a simple regression fine-tuning step -- we justify such a paradigm from a causal perspective. Following this principle, we propose a new time series representation learning framework for long sequence time series forecasting named CoST, which

applies contrastive learning methods to learn disentangled seasonal-trend representations. CoST comprises both time domain and frequency domain contrastive losses to learn discriminative trend and seasonal representations, respectively. Extensive experiments on real-world datasets show that CoST consistently outperforms the state-of-the-art methods by a considerable margin, achieving a 21.3% improvement in MSE on multivariate benchmarks. It is also robust to various choices of backbone encoders, as well as downstream regressors. Code is available at https://github.com/salesforce/CoST.

CoordX: Accelerating Implicit Neural Representation with a Split MLP Architecture

- Ruofan Liang, Hongyi Sun, Nandita Vijaykumar
- abstract@open-review(Poster): Implicit neural representations with multi-layer perceptrons (MLPs) have recently gained prominence for a wide variety of tasks such as novel view synthesis and 3D object representation and rendering. However, a significant challenge with these representations is that both training and inference with an MLP over a large number of input coordinates to learn and represent an image, video, or 3D object, require large amounts of computation and incur long processing times. In this work, we aim to accelerate inference and training of coordinate-based MLPs for implicit neural representations by proposing a new split MLP architecture, CoordX. With CoordX, the initial layers are split to learn each dimension of the input coordinates separately. The intermediate features are then fused by the last layers to generate the learned signal at the corresponding coordinate point. This significantly reduces the amount of computation required and leads to large speedups in training and inference, while achieving similar accuracy as the baseline MLP. This approach thus aims at first learning functions that are a decomposition of the original signal and then fusing them to generate the learned signal. Our proposed architecture can be generally used for many implicit neural representation tasks with no additional memory overheads. We demonstrate a speedup of up to 2.92x compared to the baseline model for image, video, and 3D shape representation and rendering tasks.

Plant 'n' Seek: Can You Find the Winning Ticket?

- Jonas Fischer, Rebekka Burkholz
- abstract@open-review(Poster): The lottery ticket hypothesis has sparked the rapid development of pruning algorithms that aim to reduce the computational costs associated with deep learning during training and model deployment. Currently, such algorithms are primarily evaluated on imaging data, for which we lack ground truth information and thus the understanding of how sparse lottery tickets could be. To fill this gap, we develop a framework that allows us to plant and hide winning tickets with desirable properties in randomly initialized neural networks. To analyze the ability of state-of-the-art pruning to identify tickets of extreme sparsity, we design and hide such tickets solving four challenging tasks. In extensive experiments, we observe similar trends as in imaging studies, indicating that our framework can provide transferable insights into realistic problems. Additionally, we can now see beyond such relative trends and highlight limitations of current pruning methods. Based on our results, we conclude that the current limitations in ticket sparsity are likely of algorithmic rather than fundamental nature. We anticipate that comparisons to planted tickets will facilitate future developments of efficient pruning algorithms.

Coherence-based Label Propagation over Time Series for Accelerated Active Learning

- Yooju Shin, Susik Yoon, Sundong Kim, Hwanjun Song, Jae-Gil Lee, Byung Suk Lee
- abstract@open-review(Poster): Time-series data are ubiquitous these days, but lack of the labels in time-series data is regarded as a hurdle for its broad applicability. Meanwhile, active learning has been successfully adopted to reduce the labeling efforts in various tasks. Thus, this paper addresses an important issue, time-series active learning. Inspired by the temporal coherence in time-series data, where consecutive data points tend to have the same label, our label propagation framework, called TCLP, automatically assigns a queried label to the data points within an accurately estimated time-series segment, thereby significantly boosting the impact of an individual query. Compared with traditional time-series active learning, TCLP is shown to improve the classification accuracy by up to 7.1 times when only 0.8% of data points in the entire time series are queried for their labels.

A Class of Short-term Recurrence Anderson Mixing Methods and Their Applications

- Fuchao Wei, Chenglong Bao, Yang Liu
- abstract@open-review(Poster): Anderson mixing (AM) is a powerful acceleration method for fixed-point iterations, but its computation requires storing many historical iterations. The extra memory footprint can be prohibitive when solving high-dimensional problems in a resource-limited machine. To reduce the memory overhead, we propose a novel class of short-term recurrence AM methods (ST-AM). The ST-AM methods only store two previous iterations with cheap corrections. We prove that the basic version of ST-AM is equivalent to the full-memory AM in strongly convex quadratic optimization, and with minor changes it has local linear convergence for solving general nonlinear fixed-point problems. We further analyze the convergence properties of the regularized ST-AM for nonconvex (stochastic) optimization. Finally, we apply ST-AM to several applications including solving root-finding problems and training neural networks. Experimental results show that ST-AM is competitive with the long-memory AM and outperforms many existing optimizers.

The Geometry of Memoryless Stochastic Policy Optimization in Infinite-Horizon POMDPs

- Johannes MÃ¼ller, Guido Montufar
- abstract@open-review(Poster): We consider the problem of finding the best memoryless stochastic policy for an infinite-horizon partially observable Markov decision process (POMDP) with finite state and action spaces with respect to either the discounted or mean reward criterion. We show that the (discounted) state-action frequencies and the expected cumulative reward are rational functions of the policy, whereby the degree is determined by the degree of partial observability. We then describe the optimization problem as a linear optimization problem in the space of feasible state-action frequencies subject to polynomial constraints that we characterize explicitly. This allows us to address the combinatorial and geometric complexity of the optimization problem using recent tools from polynomial optimization. In particular, we demonstrate how the partial observability constraints can lead to multiple smooth and non-smooth local optimizers and we estimate the number of critical points.

Efficient Sharpness-aware Minimization for Improved Training of Neural Networks

- Jiawei Du, Hanshu Yan, Jiashi Feng, Joey Tianyi Zhou, Liangli Zhen, Rick Siow Mong Goh, Vincent Tan
- abstract@open-review(Poster): Overparametrized Deep Neural Networks (DNNs) often achieve astounding performances, but may potentially result in severe generalization error. Recently, the relation between the sharpness of the loss landscape and the generalization error has been established by Foret et al. (2020), in which the Sharpness Aware Minimizer (SAM) was proposed to mitigate the degradation of the generalization. Unfortunately, SAMâ€™s computational cost is roughly double that of base optimizers, such as Stochastic Gradient Descent (SGD). This paper thus proposes Efficient Sharpness Aware Minimizer (ESAM), which boosts SAMâ€™s efficiency at no cost to its generalization performance. ESAM includes two novel and efficient training strategiesâ€”StochasticWeight Perturbation and Sharpness-Sensitive Data Selection. In the former, the sharpness measure is approximated by perturbing a stochastically chosen set of weights in each iteration; in the latter, the SAM loss is optimized using only a judiciously selected subset of data that is sensitive to the sharpness. We provide theoretical explanations as to why these strategies perform well. We also show, via extensive experiments on the CIFAR and ImageNet datasets, that ESAM enhances the efficiency over SAM from requiring 100% extra computations to 40% vis-`a-vis base optimizers, while test accuracies are preserved or even improved.

Lipschitz-constrained Unsupervised Skill Discovery

- Seohong Park, Jongwook Choi, Jaekyeom Kim, Honglak Lee, Gunhee Kim
- abstract@open-review(Poster): We study the problem of unsupervised skill discovery, whose goal is to learn a set of diverse and useful skills with no external reward. There have been a number of skill discovery methods based on maximizing the mutual information (MI) between skills and states. However, we point out that their MI objectives usually prefer static skills to dynamic ones, which may hinder the application for downstream tasks. To address this issue, we propose Lipschitz-constrained Skill Discovery (LSD), which encourages the agent to discover more diverse, dynamic, and far-reaching skills. Another benefit of LSD is that its learned representation function can be utilized for solving goal-following downstream tasks even in a zero-shot manner â€" i.e., without further training or complex planning. Through experiments on various MuJoCo robotic locomotion and manipulation environments, we demonstrate that LSD outperforms previous approaches in terms of skill diversity, state space coverage, and performance on seven downstream tasks including the challenging task of following multiple goals on Humanoid. Our code and videos are available at https://shpark.me/projects/lsd/.

Learning Generalizable Representations for Reinforcement Learning via Adaptive Meta-learner of Behavioral Similarities

- Jianda Chen, Sinno Pan
- abstract@open-review(Poster): How to learn an effective reinforcement learning-based model for control tasks from high-level visual observations is a practical and challenging problem. A key to solving this problem is to learn low-dimensional state representations from observations, from which an effective policy can be learned. In order to boost the learning of state encoding, recent works are focused on capturing behavioral similarities between state representations or applying data augmentation on visual observations. In this paper, we propose a novel meta-learner-based framework for representation learning regarding behavioral similarities for reinforcement learning. Specifically, our framework encodes the high-dimensional observations into two decomposed embeddings regarding reward and dynamics in a Markov Decision Process (MDP). A pair of meta-learners are developed, one of which quantifies the reward similarity and the other quantifies dynamics similarity over the correspondingly decomposed embeddings. The meta-learners are self-learned to update the state embeddings by approximating two disjoint terms in on-policy bisimulation metric. To incorporate the reward and dynamics terms, we further develop a strategy to adaptively balance their impacts based on different tasks or environments. We empirically demonstrate that our proposed framework outperforms state-of-the-art baselines on several benchmarks, including conventional DM Control Suite, Distracting DM Control Suite and a self-driving task CARLA.

Effective Model Sparsification by Scheduled Grow-and-Prune Methods

- Xiaolong Ma, Minghai Qin, Fei Sun, Zejiang Hou, Kun Yuan, Yi Xu, Yanzhi Wang, Yen-Kuang Chen, Rong Jin, Yuan Xie
- abstract@open-review(Poster): Deep neural networks (DNNs) are effective in solving many real-world problems. Larger DNN models usually exhibit better quality (e.g., accuracy) but their excessive computation results in long inference time. Model sparsification can reduce the computation and memory cost while maintaining model quality. Most existing sparsification algorithms unidirectionally remove weights, while others randomly or greedily explore a small subset of weights in each layer for pruning. The limitations of these algorithms reduce the level of achievable sparsity. In addition, many algorithms still require pre-trained dense models and thus suffer from large memory footprint. In this paper, we propose a novel scheduled grow-and-prune (GaP) methodology without having to pre-train a dense model. It addresses the shortcomings of the previous works by repeatedly growing a subset of layers to dense and then pruning them back to sparse after some training. Experiments show that the models pruned using the proposed methods match or beat the quality of the highly optimized dense models at 80% sparsity on a variety of tasks, such as image classification, objective detection, 3D object part segmentation, and translation. They also outperform other state-of-the-art (SOTA) methods for model sparsification. As an example, a 90% non-uniform sparse ResNet-50 model obtained via GaP achieves 77.9% top-1 accuracy on ImageNet, improving the previous SOTA results by 1.5%. Code available at: https://github.com/boone891214/GaP.

FILIP: Fine-grained Interactive Language-Image Pre-Training

- Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, Chunjing Xu
- abstract@open-review(Poster): Unsupervised large-scale vision-language pre-training has shown promising advances on various downstream tasks. Existing methods often model the cross-modal interaction either via the similarity of the global feature of each modality which misses sufficient information, or finer-grained interactions using cross/self-attention upon visual and textual tokens. However, cross/self-attention suffers from inferior efficiency in both training and inference. In this paper, we introduce a large-scale Fine-grained Interactive Language-Image Pre-training (FILIP) to achieve finer-level alignment through a cross-modal late interaction mechanism, which uses a token-wise maximum similarity between visual and textual tokens to guide the contrastive objective. FILIP successfully leverages the finer-grained expressiveness between image patches and textual words by modifying only contrastive loss, while simultaneously gaining the ability to pre-compute image and text representations offline at inference, keeping both large-scale training and inference efficient. Furthermore, we construct a new large-scale image-text pair dataset called FILIP300M for pre-training. Experiments show that FILIP achieves state-of-the-art performance on multiple downstream vision-language tasks including zero-shot image classification and image-text retrieval. The visualization on word-patch alignment further shows that FILIP can learn meaningful fine-grained features with promising localization ability.

Information Prioritization through Empowerment in Visual Model-based RL

- Homanga Bharadhwaj, Mohammad Babaeizadeh, Dumitru Erhan, Sergey Levine
- abstract@open-review(Poster): Model-based reinforcement learning (RL) algorithms designed for handling complex visual observations typically learn some sort of latent state representation, either explicitly or implicitly. Standard methods of this sort do not distinguish between functionally relevant aspects of the state and irrelevant distractors, instead aiming to represent all available information equally. We propose a modified objective for model-based RL that, in combination with mutual information maximization, allows us to learn representations and dynamics for visual model-based RL without reconstruction in a way that explicitly prioritizes functionally relevant factors. The key principle behind our design is to integrate a term inspired by variational empowerment into a state-space learning model based on mutual information. This term prioritizes information that is correlated with action, thus ensuring that functionally relevant factors are captured first. Furthermore, the same empowerment term also promotes faster exploration during the RL process, especially for sparse-reward tasks where the reward signal is insufficient to drive exploration in the early stages of learning. We evaluate the approach on a suite of vision-based robot control tasks with natural video backgrounds, and show that the proposed prioritized information objective outperforms state-of-the-art model based RL approaches by an average of 20\% in terms of episodic returns at 1M environment interactions with 30\% higher sample efficiency at 100k interactions.

Efficient Active Search for Combinatorial Optimization Problems

- AndrÃ© Hottung, Yeong-Dae Kwon, Kevin Tierney
- abstract@open-review(Poster): Recently numerous machine learning based methods for combinatorial optimization problems have been proposed that learn to construct solutions in a sequential decision process via reinforcement learning. While these methods can be easily combined with search strategies like sampling and beam search, it is not straightforward to integrate them into a high-level search procedure offering strong search guidance. Bello et al. (2016) propose active search, which adjusts the weights of a (trained) model with respect to a single instance at test time using reinforcement learning. While active search is simple to implement, it is not competitive with state-of-the-art methods because adjusting all model weights for each test instance is very time and memory intensive. Instead of updating all model weights, we propose and evaluate three efficient active search strategies that only update a subset of parameters during the search. The proposed methods offer a simple way to significantly improve the search performance of a given model and outperform state-of-the-art machine learning based methods on combinatorial problems, even surpassing the well-known heuristic solver LKH3 on the

capacitated vehicle routing problem. Finally, we show that (efficient) active search enables learned models to effectively solve instances that are much larger than those seen during training.

Ancestral protein sequence reconstruction using a tree-structured Ornstein-Uhlenbeck variational autoencoder

- Lys Sanz Moreta, Ola RÃ¸nning, Ahmad Salim Al-Sibahi, Jotun Hein, Douglas Theobald, Thomas Hamelryck
- abstract@open-review(Poster): We introduce a deep generative model for representation learning of biological sequences that, unlike existing models, explicitly represents the evolutionary process. The model makes use of a tree-structured Ornstein-Uhlenbeck process, obtained from a given phylogenetic tree, as an informative prior for a variational autoencoder. We show the model performs well on the task of ancestral sequence reconstruction of single protein families. Our results and ablation studies indicate that the explicit representation of evolution using a suitable tree-structured prior has the potential to improve representation learning of biological sequences considerably. Finally, we briefly discuss extensions of the model to genomic-scale data sets and the case of a latent phylogenetic tree.

Training Structured Neural Networks Through Manifold Identification and Variance Reduction

- Zih-Syuan Huang, Ching-pei Lee
- abstract@open-review(Poster): This paper proposes an algorithm, RMDA, for training neural networks (NNs) with a regularization term for promoting desired structures. RMDA does not incur computation additional to proximal SGD with momentum, and achieves variance reduction without requiring the objective function to be of the finite-sum form. Through the tool of manifold identification from nonlinear optimization, we prove that after a finite number of iterations, all iterates of RMDA possess a desired structure identical to that induced by the regularizer at the stationary point of asymptotic convergence, even in the presence of engineering tricks like data augmentation that complicate the training process. Experiments on training NNs with structured sparsity confirm that variance reduction is necessary for such an identification, and show that RMDA thus significantly outperforms existing methods for this task. For unstructured sparsity, RMDA also outperforms a state-of-the-art pruning method, validating the benefits of training structured NNs through regularization. Implementation of RMDA is available at https://www.github.com/zihsyuan1214/rmda.

The Neural Data Router: Adaptive Control Flow in Transformers Improves Systematic Generalization

- RÃ³bert CsordÃ¡s, Kazuki Irie, JÃ¼rgen Schmidhuber
- abstract@open-review(Poster): Despite progress across a broad range of applications, Transformers have limited success in systematic generalization. The situation is especially frustrating in the case of algorithmic tasks, where they often fail to find intuitive solutions that route relevant information to the right node/operation at the right time in the grid represented by Transformer columns. To facilitate the learning of useful control flow, we propose two modifications to the Transformer architecture, copy gate and geometric attention. Our novel Neural Data Router (NDR) achieves 100% length generalization accuracy on the classic compositional table lookup task, as well as near-perfect accuracy on the simple arithmetic task and a new variant of ListOps testing for generalization across computational depths. NDRâ€™s attention and gating patterns tend to be interpretable as an intuitive form of neural routing

On the Limitations of Multimodal VAEs

- Imant Daunhawer, Thomas M. Sutter, Kieran Chin-Cheong, Emanuele Palumbo, Julia E Vogt
- abstract@open-review(Poster): Multimodal variational autoencoders (VAEs) have shown promise as efficient generative models for weakly-supervised data. Yet, despite their advantage of weak supervision, they exhibit a gap in generative quality compared to unimodal VAEs, which are completely unsupervised. In an attempt to explain this gap, we uncover a fundamental limitation that applies to a large family of mixture-based multimodal VAEs. We prove that the sub-sampling of modalities enforces an undesirable upper bound on the multimodal ELBO and thereby limits the generative quality of the respective models. Empirically, we showcase the generative quality gap on both synthetic and real data and present the tradeoffs between different variants of multimodal VAEs. We find that none of the existing approaches fulfills all desired criteria of an effective multimodal generative model when applied on more complex datasets than those used in previous benchmarks. In summary, we identify, formalize, and validate fundamental limitations of VAE-based approaches for modeling weakly-supervised data and discuss implications for real-world applications.

Recursive Disentanglement Network

- Yixuan Chen, Yubin Shi, Dongsheng Li, Yujiang Wang, Mingzhi Dong, Yingying Zhao, Robert Dick, Qin Lv, Fan Yang, Li Shang
- abstract@open-review(Poster): Disentangled feature representation is essential for data-efficient learning. The feature space of deep models is inherently compositional. Existing β-VAE-based methods, which only apply disentanglement regularization to the resulting embedding space of deep models, cannot effectively regularize such compositional feature space, resulting in unsatisfactory disentangled results. In this paper, we formulate the compositional disentanglement learning problem from an information-theoretic perspective and propose a recursive disentanglement network (RecurD) that propagates regulatory inductive bias recursively across the compositional feature space during disentangled representation learning. Experimental studies demonstrate that RecurD outperforms β-VAE and several of its state-of-the-art variants on disentangled representation learning and enables more data-efficient downstream machine learning tasks.

ADAVI: Automatic Dual Amortized Variational Inference Applied To Pyramidal Bayesian Models

- Louis Rouillard, Demian Wassermann
- abstract@open-review(Poster): Frequently, population studies feature pyramidally-organized data represented using Hierarchical Bayesian Models (HBM) enriched with plates. These models can become prohibitively large in settings such as neuroimaging, where a sample is composed of a functional MRI signal measured on 300 brain locations, across 4 measurement sessions, and 30 subjects, resulting in around 1 million latent parameters.

Such high dimensionality hampers the usage of modern, expressive flow-based techniques.

To infer parameter posterior distributions in this challenging class of problems, we designed a novel methodology that automatically produces a variational family dual to a target HBM. This variational family, represented as a neural network, consists in the combination of an attention-based hierarchical encoder feeding summary statistics to a set of normalizing flows. Our automatically-derived neural network exploits exchangeability in the plate-enriched HBM and factorizes its parameter space. The resulting architecture reduces by orders of magnitude its parameterization with respect to that of a typical flow-based representation, while maintaining expressivity.

Our method performs inference on the specified HBM in an amortized setup: once trained, it can readily be applied to a new data sample to compute the parameters' full posterior.

We demonstrate the capability and scalability of our method on simulated data, as well as a challenging high-dimensional brain parcellation experiment. We also open up several questions that lie at the intersection between normalizing flows, SBI, structured Variational Inference, and inference amortization.

Distributionally Robust Models with Parametric Likelihood Ratios

- Paul Michel, Tatsunori Hashimoto, Graham Neubig
- abstract@open-review(Poster): As machine learning models are deployed ever more broadly, it becomes increasingly important that they are not only able to perform well on their training distribution, but also yield accurate predictions when confronted with distribution shift. The Distributionally Robust Optimization (DRO) framework proposes to address this issue by training models to minimize their expected risk under a collection of distributions, to imitate test-time shifts. This is most commonly achieved by instance-level re-weighting of the training objective to emulate the likelihood ratio with possible test distributions, which allows for estimating their empirical risk via importance sampling (assuming that they are subpopulations of the training distribution). However, re-weighting schemes in the literature are usually limited due to the difficulty of keeping the optimization problem tractable and the complexity of enforcing normalization constraints. In this paper, we show that three simple ideas -- mini-batch level normalization, a KL penalty and simultaneous gradient updates -- allow us to train models with DRO using a broader class of parametric likelihood ratios. In a series of experiments on both image and text classification benchmarks, we find that models trained with the resulting parametric adversaries are consistently more robust to subpopulation shifts when compared to other DRO approaches, and that the method performs reliably well with little hyper-parameter tuning.

Constrained Physical-Statistics Models for Dynamical System Identification and Prediction

- JÃ©rÃ©mie DONA, Marie DÃ©chelle, patrick gallinari, Marina Levy
- abstract@open-review(Poster): Modeling dynamical systems combining prior physical knowledge and machine learning (ML) is promising in scientific problems when the underlying processes are not fully understood, e.g. when the dynamics is partially known. A common practice to identify the respective parameters of the physical and ML components is to formulate the problem as supervised learning on observed trajectories. However, this formulation leads to an infinite number of possible decompositions. To solve this ill-posedness, we reformulate the learning problem by introducing an upper bound on the prediction error of a physical-statistical model. This allows us to control the contribution of both the physical and statistical components to the overall prediction. This framework generalizes several existing hybrid schemes proposed in the literature. We provide theoretical guarantees on the well-posedness of our formulation along with a proof of convergence in a simple affine setting. For more complex dynamics, we validate our framework experimentally.

Doubly Adaptive Scaled Algorithm for Machine Learning Using Second-Order Information

- Majid Jahani, Sergey Rusakov, Zheng Shi, Peter RichtÃ¡rik, Michael W. Mahoney, Martin Takac
- abstract@open-review(Poster): We present a novel adaptive optimization algorithm for large-scale machine learning problems. Equipped with a low-cost estimate of local curvature and Lipschitz smoothness, our method dynamically adapts the search direction and step-size. The search direction contains gradient information preconditioned by a well-scaled diagonal preconditioning matrix that captures the local curvature information. Our methodology does not require the tedious task of learning rate tuning, as the learning rate is updated automatically without adding an extra hyper-parameter. We provide convergence guarantees on a comprehensive collection of optimization problems, including convex, strongly convex, and nonconvex problems, in both deterministic and stochastic regimes. We also conduct an extensive empirical evaluation on standard machine learning problems, justifying our algorithm's versatility and demonstrating its strong performance compared to other start-of-the-art first-order and second-order methods.

Understanding approximate and unrolled dictionary learning for pattern recovery

- BenoÃ®t MalÃ©zieux, Thomas Moreau, Matthieu Kowalski
- abstract@open-review(Poster): Dictionary learning consists of finding a sparse representation from noisy data and is a common way to encode data-driven prior knowledge on signals. Alternating minimization (AM) is standard for the underlying optimization, where gradient descent steps alternate with sparse coding procedures. The major drawback of this method is its prohibitive computational cost, making it unpractical on large real-world data sets. This work studies an approximate formulation of dictionary learning based on unrolling and compares it to alternating minimization to find the best trade-off between speed and precision. We analyze the asymptotic behavior and convergence rate of gradients estimates in both methods. We show that unrolling performs better on the support of the inner problem solution and during the first iterations. Finally, we apply unrolling on pattern learning in magnetoencephalography (MEG) with the help of a stochastic algorithm and compare the performance to a state-of-the-art method.

Constraining Linear-chain CRFs to Regular Languages

- Sean Papay, Roman Klinger, Sebastian Pado
- abstract@open-review(Poster): A major challenge in structured prediction is to represent the interdependencies within output structures. When outputs are structured as sequences, linear-chain conditional random fields (CRFs) are a widely used model class which can learn local dependencies in the output. However, the CRF's Markov assumption makes it impossible for CRFs to represent distributions with nonlocal dependencies, and standard CRFs are unable to respect nonlocal constraints of the data (such as global arity constraints on output labels). We present a generalization of CRFs that can enforce a broad class of constraints, including nonlocal ones, by specifying the space of possible output structures as a regular language $\mathcal{L}$. The resulting regular-constrained CRF (RegCCRF) has the same formal properties as a standard CRF, but assigns zero probability to all label sequences not in $\mathcal{L}$. Notably, RegCCRFs can incorporate their constraints during training, while related models only enforce constraints during decoding. We prove that constrained training is never worse than constrained decoding, and show empirically that it can be substantially better in practice. Additionally, we demonstrate a practical benefit on downstream tasks by incorporating a RegCCRF into a deep neural model for semantic role labeling, exceeding state-of-the-art results on a standard dataset.

Dive Deeper Into Integral Pose Regression

- Kerui Gu, Linlin Yang, Angela Yao
- abstract@open-review(Poster): Integral pose regression combines an implicit heatmap with end-to-end training for human body and hand pose estimation. Unlike detection-based heatmap methods, which decode final joint positions from the heatmap with a non-differentiable argmax operation, integral regression methods apply a differentiable expectation operation. This paper offers a deep dive into the inference and back-propagation of integral pose regression to better understand the differences in performance and training compared to detection-based methods. For inference, we give theoretical support as to why expectation should always be better than the argmax operation, i.e. integral regression should always outperform detection. Yet, in practice, this is observed only in hard cases because the heatmap activation for regression shrinks in easy cases. We then experimentally show that activation shrinkage is one of the leading causes for integral regression's inferior performance. For back-propagation, we theoretically and empirically analyze the gradients to explain the slow training speed of integral regression. Based on these findings, we incorporate the supervision of a spatial prior to speed up training and improve performance.

Evidential Turing Processes

- Melih Kandemir, Abdullah AkgÃ¼l, Manuel Haussmann, Gozde Unal
- abstract@open-review(Poster): A probabilistic classifier with reliable predictive uncertainties i) fits successfully to the target domain data, ii) provides calibrated class probabilities in difficult regions of the target domain (e.g. class overlap), and iii) accurately identifies queries coming out of the target domain and reject them. We introduce an original combination of Evidential Deep Learning, Neural Processes, and Neural Turing Machines capable of providing all three essential properties mentioned above for total uncertainty quantification. We observe our method on three image classification benchmarks to consistently improve the in-domain uncertainty quantification, out-of-domain detection, and robustness against input perturbations with

one single model. Our unified solution delivers an implementation-friendly and computationally efficient recipe for safety clearance and provides intellectual economy to an investigation of algorithmic roots of epistemic awareness in deep neural nets.

Noisy Feature Mixup

- Soon Hoe Lim, N. Benjamin Erichson, Francisco Utrera, Winnie Xu, Michael W. Mahoney
- abstract@open-review(Poster): We introduce Noisy Feature Mixup (NFM), an inexpensive yet effective method for data augmentation that combines the best of interpolation based training and noise injection schemes. Rather than training with convex combinations of pairs of examples and their labels, we use noise-perturbed convex combinations of pairs of data points in both input and feature space. This method includes mixup and manifold mixup as special cases, but it has additional advantages, including better smoothing of decision boundaries and enabling improved model robustness. We provide theory to understand this as well as the implicit regularization effects of NFM. Our theory is supported by empirical results, demonstrating the advantage of NFM, as compared to mixup and manifold mixup. We show that residual networks and vision transformers trained with NFM have favorable trade-offs between predictive accuracy on clean data and robustness with respect to various types of data perturbation across a range of computer vision benchmark datasets.

Peek-a-Boo: What (More) is Disguised in a Randomly Weighted Neural Network, and How to Find It Efficiently

- Xiaohan Chen, Jason Zhang, Zhangyang Wang
- abstract@open-review(Poster): Sparse neural networks (NNs) are intensively investigated in literature due to their appeal in saving storage, memory, and computational costs. A recent work (Ramanujan et al., 2020) showed that, different from conventional pruning-and-finetuning pipeline, there exist hidden subnetworks in randomly initialized NNs that have good performance without training the weights. However, such "hidden subnetworks" have mediocre performances and require an expensive edge-popup algorithm to search for them. In this work, we define an extended class of subnetworks in randomly initialized NNs called disguised subnetworks, which are not only "hidden" in the random networks but also "disguised" -- hence can only be "unmasked" with certain transformations on weights. We argue that the unmasking process plays an important role in enlarging the capacity of the subnetworks and thus grants two major benefits: (i) the disguised subnetworks easily outperform the hidden counterparts; (ii) the unmasking process helps to relax the quality requirement on the sparse subnetwork mask so that the expensive edge-popup algorithm can be replaced with more efficient alternatives. On top of this new concept, we propose a novel two-stage algorithm that plays a Peek-a-Boo (PaB) game to identify the disguised subnetworks with a combination of two operations: (1) searching efficiently for a subnetwork at random initialization; (2) unmasking the disguise by learning to transform the resulting subnetwork's remaining weights. Furthermore, we show that the unmasking process can be efficiently implemented (a) without referring to any latent weights or scores; and (b) by only leveraging approximated gradients, so that the whole training algorithm is computationally light. Extensive experiments with several large models (ResNet-18, ResNet-50, and WideResNet-28) and datasets (CIFAR-10, CIFAR-100 and ImageNet) demonstrate the competency of PaB over edge-popup and other counterparts. Our codes are available at: https://github.com/VITA-Group/Peek-a-Boo.

How Well Does Self-Supervised Pre-Training Perform with Streaming Data?

- Dapeng Hu, Shipeng Yan, Qizhengqiu Lu, Lanqing HONG, Hailin Hu, Yifan Zhang, Zhenguo Li, Xinchao Wang, Jiashi Feng
- abstract@open-review(Poster): Prior works on self-supervised pre-training focus on the joint training scenario, where massive unlabeled data are assumed to be given as input all at once, and only then is a learner trained. Unfortunately, such a problem setting is often impractical if not infeasible since many real-world tasks rely on sequential learning, e.g., data are decentralized or collected in a streaming fashion. In this paper, we conduct the first thorough and dedicated investigation on self-supervised pre-training with streaming data, aiming to shed light on the model behavior under this overlooked setup. Specifically, we pre-train over 500 models on four categories of pre-training streaming data from ImageNet and DomainNet and evaluate them on three types of downstream tasks and 12 different downstream datasets. Our studies show that, somehow beyond our expectation, with simple data replay or parameter regularization, sequential self-supervised pre-training turns out to be an efficient alternative for joint pre-training, as the performances of the former are mostly on par with those of the latter. Moreover, catastrophic forgetting, a common issue in sequential supervised learning, is much alleviated in sequential self-supervised learning (SSL), which is well justified through our comprehensive empirical analysis on representations and the sharpness of minima in the loss landscape. Our findings, therefore, suggest that, in practice, for SSL, the cumbersome joint training can be replaced mainly by sequential learning, which in turn enables a much broader spectrum of potential application scenarios.

Subspace Regularizers for Few-Shot Class Incremental Learning

- Afra Feyza AkyÃ¼rek, Ekin AkyÃ¼rek, Derry Wijaya, Jacob Andreas
- abstract@open-review(Poster): Few-shot class incremental learning---the problem of updating a trained classifier to discriminate among an expanded set of classes with limited labeled data---is a key challenge for machine learning systems deployed in non-stationary environments. Existing approaches to the problem rely on complex model architectures and training procedures that are difficult to tune and re-use. In this paper, we present an extremely simple approach that enables the use of ordinary logistic regression classifiers for few-shot incremental learning. The key to this approach is a new family of \textit{subspace regularization} schemes that encourage weight vectors for new classes to lie close to the subspace spanned by the weights of existing classes. When combined with pretrained convolutional feature extractors, logistic regression models trained with subspace regularization outperform specialized, state-of-the-art approaches to few-shot incremental image classification by up to 23\% on the \textit{mini}ImageNet dataset. Because of its simplicity, subspace regularization can be straightforwardly configured to incorporate additional background information about the new classes (including class names and descriptions specified in natural language); this offers additional control over the trade-off between existing and new classes. Our results show that simple geometric regularization of class representations offers an effective tool for continual learning.

Using Graph Representation Learning with Schema Encoders to Measure the Severity of Depressive Symptoms

- Simin Hong, Anthony Cohn, David Crossland Hogg
- abstract@open-review(Poster): Graph neural networks (GNNs) are widely used in regression and classification problems applied to text, in areas such as sentiment analysis and medical decision-making processes. We propose a novel form for node attributes within a GNN based model that captures node-specific embeddings for every word in the vocabulary. This provides a global representation at each node, coupled with node-level updates according to associations among words in a transcript. We demonstrate the efficacy of the approach by augmenting the accuracy of measuring major depressive disorder (MDD). Prior research has sought to make a diagnostic prediction of depression levels from patient data using several modalities, including audio, video, and text. On the DAIC-WOZ benchmark, our method outperforms state-of-art methods by a substantial margin, including those using multiple modalities. Moreover, we also evaluate the performance of our novel model on a Twitter sentiment dataset. We show that our model outperforms a general GNN model by leveraging our novel 2-D node attributes. These results demonstrate the generality of the proposed method.

Actor-Critic Policy Optimization in a Large-Scale Imperfect-Information Game

- Haobo Fu, Weiming Liu, Shuang Wu, Yijia Wang, Tao Yang, Kai Li, Junliang Xing, Bin Li, Bo Ma, QIANG FU, Yang Wei
- abstract@open-review(Poster): The deep policy gradient method has demonstrated promising results in many large-scale games, where the agent learns purely from its own experience. Yet, policy gradient methods with self-play suffer convergence problems to a Nash Equilibrium (NE) in multi-agent situations. Counterfactual regret minimization (CFR) has a convergence guarantee to a NE in 2-player zero-sum games, but it usually needs domain-specific abstractions to deal with large-scale games. Inheriting merits from both methods, in this paper we extend the actor-critic algorithm framework in deep reinforcement learning to tackle a large-scale 2-player zero-sum imperfect-information game, 1-on-1 Mahjong, whose information set size and game

length are much larger than poker. The proposed algorithm, named Actor-Critic Hedge (ACH), modifies the policy optimization objective from originally maximizing the discounted returns to minimizing a type of weighted cumulative counterfactual regret. This modification is achieved by approximating the regret via a deep neural network and minimizing the regret via generating self-play policies using Hedge. ACH is theoretically justified as it is derived from a neural-based weighted CFR, for which we prove the convergence to a NE under certain conditions. Experimental results on the proposed 1-on-1 Mahjong benchmark and benchmarks from the literature demonstrate that ACH outperforms related state-of-the-art methods. Also, the agent obtained by ACH defeats a human champion in 1-on-1 Mahjong.

Policy Gradients Incorporating the Future

- David Venuto, Elaine Lau, Doina Precup, Ofir Nachum
- abstract@open-review(Poster): Reasoning about the future -- understanding how decisions in the present time affect outcomes in the future -- is one of the central challenges for reinforcement learning (RL), especially in highly-stochastic or partially observable environments. While predicting the future directly is hard, in this work we introduce a method that allows an agent to ``look into the future" without explicitly predicting it. Namely, we propose to allow an agent, during its training on past experience, to observe what \emph{actually} happened in the future at that time, while enforcing an information bottleneck to avoid the agent overly relying on this privileged information. Coupled with recent advances in variational inference and a latent-variable autoregressive model, this gives our agent the ability to utilize rich and \emph{useful} information about the future trajectory dynamics in addition to the present. Our method, Policy Gradients Incorporating the Future (PGIF), is easy to implement and versatile, being applicable to virtually any policy gradient algorithm. We apply our proposed method to a number of off-the-shelf RL algorithms and show that PGIF is able to achieve higher reward faster in a variety of online and offline RL domains, as well as sparse-reward and partially observable environments.

Gradient Information Matters in Policy Optimization by Back-propagating through Model

- Chongchong Li, Yue Wang, Wei Chen, Yuting Liu, Zhi-Ming Ma, Tie-Yan Liu
- abstract@open-review(Poster): Model-based reinforcement learning provides an efficient mechanism to find the optimal policy by interacting with the learned environment. In addition to treating the learned environment like a black-box simulator, a more effective way to use the model is to exploit its differentiability. Such methods require the gradient information of the learned environment model when calculating the policy gradient. However, since the error of gradient is not considered in the model learning phase, there is no guarantee for the model's accuracy. To address this problem, we first analyze the convergence rate for the policy optimization methods when the policy gradient is calculated using the learned environment model. The theoretical results show that the model gradient error matters in the policy optimization phrase. Then we propose a two-model-based learning method to control the prediction error and the gradient error. We separate the different roles of these two models at the model learning phase and coordinate them at the policy optimization phase. After proposing the method, we introduce the directional derivative projection policy optimization (DDPPO) algorithm as a practical implementation to find the optimal policy. Finally, we empirically demonstrate the proposed algorithm has better sample efficiency when achieving a comparable or better performance on benchmark continuous control tasks.

VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning

- Adrien Bardes, Jean Ponce, Yann LeCun
- abstract@open-review(Poster): Recent self-supervised methods for image representation learning maximize the agreement between embedding vectors produced by encoders fed with different views of the same image. The main challenge is to prevent a collapse in which the encoders produce constant or non-informative vectors. We introduce VICReg (Variance-Invariance-Covariance Regularization), a method that explicitly avoids the collapse problem with two regularizations terms applied to both embeddings separately: (1) a term that maintains the variance of each embedding dimension above a threshold, (2) a term that decorrelates each pair of variables. Unlike most other approaches to the same problem, VICReg does not require techniques such as: weight sharing between the branches, batch normalization, feature-wise normalization, output quantization, stop gradient, memory banks, etc., and achieves results on par with the state of the art on several downstream tasks. In addition, we show that our variance regularization term stabilizes the training of other methods and leads to performance improvements.

High Probability Generalization Bounds with Fast Rates for Minimax Problems

- Shaojie Li, Yong Liu
- abstract@open-review(Poster): Minimax problems are receiving an increasing amount of attention in a wide range of applications in machine learning (ML), for instance, reinforcement learning, robust optimization, adversarial learning, and distributed computing, to mention but a few. Current studies focus on the fundamental understanding of general minimax problems with an emphasis on convergence behavior. As a comparison, there is far less work to study the generalization performance. Additionally, existing generalization bounds are almost all derived in expectation, and the high probability bounds are all presented in the slow order $\mathcal{O}(1/\sqrt{n})$, where n is the sample size. In this paper, we provide improved generalization analyses and obtain sharper high probability generalization bounds for most existing generalization measures of minimax problems. We then use the improved learning bounds to establish high probability generalization bounds with fast rates for classical empirical saddle point (ESP) solution and several popular gradient-based optimization algorithms, including gradient descent ascent (GDA), stochastic gradient descent ascent (SGDA), proximal point method (PPM), extra-gradient (EG), and optimistic gradient descent ascent (OGDA). In summary, we provide a systematical analysis of sharper generalization bounds of minimax problems.

SUMNAS: Supernet with Unbiased Meta-Features for Neural Architecture Search

- Hyeonmin Ha, Ji-Hoon Kim, Semin Park, Byung-Gon Chun
- abstract@open-review(Poster): One-shot Neural Architecture Search (NAS) usually constructs an over-parameterized network, which we call a supernet, and typically adopts sharing parameters among the sub-models to improve computational efficiency. One-shot NAS often repeatedly samples sub-models from the supernet and trains them to optimize the shared parameters. However, this training strategy suffers from multi-model forgetting. Training a sampled sub-model overrides the previous knowledge learned by the other sub-models, resulting in an unfair performance evaluation between the sub-models. We propose Supernet with Unbiased Meta-Features for Neural Architecture Search (SUMNAS), a supernet learning strategy based on meta-learning to tackle the knowledge forgetting issue. During the training phase, we explicitly address the multi-model forgetting problem and help the supernet learn unbiased meta-features, independent from the sampled sub-models. Once training is over, sub-models can be instantly compared to get the overall ranking or the best sub-model. Our evaluation on the NAS-Bench-201 and MobileNet-based search space demonstrate that SUMNAS shows improved ranking ability and finds architectures whose performance is on par with existing state-of-the-art NAS algorithms.

Temporal Efficient Training of Spiking Neural Network via Gradient Re-weighting

- Shikuang Deng, Yuhang Li, Shanghang Zhang, Shi Gu
- abstract@open-review(Poster): Recently, brain-inspired spiking neuron networks (SNNs) have attracted widespread research interest because of their event-driven and energy-efficient characteristics. It is difficult to efficiently train deep SNNs due to the non-differentiability of its activation function, which disables the typically used gradient descent approaches for traditional artificial neural networks (ANNs). Although the adoption of surrogate gradient (SG) formally allows for the back-propagation of losses, the discrete spiking mechanism actually differentiates the loss landscape of SNNs from that of ANNs, failing the surrogate gradient methods to achieve comparable accuracy as for ANNs. In this paper, we first analyze why the current direct training approach with surrogate gradient results in SNNs with poor generalizability. Then we introduce the temporal efficient training (TET) approach to

compensate for the loss of momentum in the gradient descent with SG so that the training process can converge into flatter minima with better generalizability. Meanwhile, we demonstrate that TET improves the temporal scalability of SNN and induces a temporal inheritable training for acceleration. Our method consistently outperforms the SOTA on all reported mainstream datasets, including CIFAR-10/100 and ImageNet. Remarkably on DVS-CIFAR10, we obtained 83% top-1 accuracy, over 10% improvement compared to existing state of the art.

Reliable Adversarial Distillation with Unreliable Teachers

- Jianing Zhu, Jiangchao Yao, Bo Han, Jingfeng Zhang, Tongliang Liu, Gang Niu, Jingren Zhou, Jianliang Xu, Hongxia Yang
- abstract@open-review(Poster): In ordinary distillation, student networks are trained with soft labels (SLs) given by pretrained teacher networks, and students are expected to improve upon teachers since SLs are stronger supervision than the original hard labels. However, when considering adversarial robustness, teachers may become unreliable and adversarial distillation may not work: teachers are pretrained on their own adversarial data, and it is too demanding to require that teachers are also good at every adversarial data queried by students. Therefore, in this paper, we propose reliable introspective adversarial distillation (IAD) where students partially instead of fully trust their teachers. Specifically, IAD distinguishes between three cases given a query of a natural data (ND) and the corresponding adversarial data (AD): (a) if a teacher is good at AD, its SL is fully trusted; (b) if a teacher is good at ND but not AD, its SL is partially trusted and the student also takes its own SL into account; (c) otherwise, the student only relies on its own SL. Experiments demonstrate the effectiveness of IAD for improving upon teachers in terms of adversarial robustness.

Neural Program Synthesis with Query

- Di Huang, Rui Zhang, Xing Hu, Xishan Zhang, Pengwei Jin, Nan Li, Zidong Du, Qi Guo, Yunji Chen
- abstract@open-review(Poster): Aiming to find a program satisfying the user intent given input-output examples, program synthesis has attracted increasing interest in the area of machine learning. Despite the promising performance of existing methods, most of their success comes from the privileged information of well-designed input-output examples. However, providing such input-output examples is unrealistic because it requires the users to have the ability to describe the underlying program with a few input-output examples under the training distribution. In this work, we propose a query-based framework that trains a query neural network to generate informative input-output examples automatically and interactively from a large query space. The quality of the query depends on the amount of the mutual information between the query and the corresponding program, which can guide the optimization of the query framework. To estimate the mutual information more accurately, we introduce the functional space (F-space) which models the relevance between the input-output examples and the programs in a differentiable way. We evaluate the effectiveness and generalization of the proposed query-based framework on the Karel task and the list processing task. Experimental results show that the query-based framework can generate informative input-output examples which achieve and even outperform well-designed input-output examples.

Delaunay Component Analysis for Evaluation of Data Representations

- Petra Poklukar, Vladislav Polianskii, Anastasiia Varava, Florian T. Pokorny, Danica Kragic Jensfelt
- abstract@open-review(Poster): Advanced representation learning techniques require reliable and general evaluation methods. Recently, several algorithms based on the common idea of geometric and topological analysis of a manifold approximated from the learned data representations have been proposed. In this work, we introduce Delaunay Component Analysis (DCA) -- an evaluation algorithm which approximates the data manifold using a more suitable neighbourhood graph called Delaunay graph. This provides a reliable manifold estimation even for challenging geometric arrangements of representations such as clusters with varying shape and density as well as outliers, which is where existing methods often fail. Furthermore, we exploit the nature of Delaunay graphs and introduce a framework for assessing the quality of individual novel data representations. We experimentally validate the proposed DCA method on representations obtained from neural networks trained with contrastive objective, supervised and generative models, and demonstrate various use cases of our extended single point evaluation framework.

Visual hyperacuity with moving sensor and recurrent neural computations

- Alexander Rivkind, Or Ram, Eldad Assa, Michael Kreiserman, Ehud Ahissar
- abstract@open-review(Poster): Dynamical phenomena, such as recurrent neuronal activity and perpetual motion of the eye, are typically overlooked in models of bottom-up visual perception. Recent experiments suggest that tiny inter-saccadic eye motion ("fixational drift") enhances visual acuity beyond the limit imposed by the density of retinal photoreceptors. Here we hypothesize that such an enhancement is enabled by recurrent neuronal computations in early visual areas. Specifically, we explore a setting involving a low-resolution dynamical sensor that moves with respect to a static scene, with drift-like tiny steps. This setting mimics a dynamical eye viewing objects in perceptually-challenging conditions. The dynamical sensory input is classified by a convolutional neural network with recurrent connectivity added to its lower layers, in analogy to recurrent connectivity in early visual areas. Applying our system to CIFAR-10 and CIFAR-100 datasets down-sampled via 8x8 sensor, we found that (i) classification accuracy, which is drastically reduced by this down-sampling, is mostly restored to its 32x32 baseline level when using a moving sensor and recurrent connectivity, (ii) in this setting, neurons in the early layers exhibit a wide repertoire of selectivity patterns, spanning the spatiotemporal selectivity space, with neurons preferring different combinations of spatial and temporal patterning, and (iii) curved sensor's trajectories improve visual acuity compared to straight trajectories, echoing recent experimental findings involving eye-tracking in challenging conditions. Our work sheds light on the possible role of recurrent connectivity in early vision as well as the roles of fixational drift and temporal-frequency selective cells in the visual system. It also proposes a solution for artificial image recognition in settings with limited resolution and multiple time samples, such as in edge AI applications.

Partial Wasserstein Adversarial Network for Non-rigid Point Set Registration

- Ziming Wang, Nan Xue, Ling Lei, Gui-Song Xia
- abstract@open-review(Poster): Given two point sets, the problem of registration is to recover a transformation that matches one set to the other. This task is challenging due to the presence of large number of outliers, the unknown non-rigid deformations and the large sizes of point sets. To obtain strong robustness against outliers, we formulate the registration problem as a partial distribution matching (PDM) problem, where the goal is to partially match the distributions represented by point sets in a metric space. To handle large point sets, we propose a scalable PDM algorithm by utilizing the efficient partial Wasserstein-1 (PW) discrepancy. Specifically, we derive the Kantorovich-Rubinstein duality for the PW discrepancy, and show its gradient can be explicitly computed. Based on these results, we propose a partial Wasserstein adversarial network (PWAN), which is able to approximate the PW discrepancy by a neural network, and minimize it by gradient descent. In addition, it also incorporates an efficient coherence regularizer for non-rigid transformations to avoid unrealistic deformations. We evaluate PWAN on practical point set registration tasks, and show that the proposed PWAN is robust, scalable and performs more favorably than the state-of-the-art methods.

Quantitative Performance Assessment of CNN Units via Topological Entropy Calculation

- Yang Zhao, Hao Zhang
- abstract@open-review(Poster): Identifying the status of individual network units is critical for understanding the mechanism of convolutional neural networks (CNNs). However, it is still challenging to reliably give a general indication of unit status, especially for units in different network models. To this end, we propose a novel method for quantitatively clarifying the status of single unit in CNN using algebraic topological tools. Unit status is indicated via the calculation of a defined topological-based entropy, called feature entropy, which measures the degree of chaos of the global spatial pattern hidden in the unit for a category. In this way, feature entropy could provide an accurate indication of status for units in different networks with diverse situations like weight-rescaling operation. Further, we show that feature entropy decreases as the layer goes deeper and shares almost simultaneous trend with loss

during training. We show that by investigating the feature entropy of units on only training data, it could give discrimination between networks with different generalization ability from the view of the effectiveness of feature representations.

Imitation Learning by Reinforcement Learning

- Kamil Ciosek
- abstract@open-review(Poster): Imitation learning algorithms learn a policy from demonstrations of expert behavior. We show that, for deterministic experts, imitation learning can be done by reduction to reinforcement learning with a stationary reward. Our theoretical analysis both certifies the recovery of expert reward and bounds the total variation distance between the expert and the imitation learner, showing a link to adversarial imitation learning. We conduct experiments which confirm that our reduction works well in practice for continuous control tasks.

On-Policy Model Errors in Reinforcement Learning

- Lukas Froehlich, Maksym Lefarov, Melanie Zeilinger, Felix Berkenkamp
- abstract@open-review(Poster): Model-free reinforcement learning algorithms can compute policy gradients given sampled environment transitions, but require large amounts of data. In contrast, model-based methods can use the learned model to generate new data, but model errors and bias can render learning unstable or suboptimal. In this paper, we present a novel method that combines real-world data and a learned model in order to get the best of both worlds. The core idea is to exploit the real-world data for on-policy predictions and use the learned model only to generalize to different actions. Specifically, we use the data as time-dependent on-policy correction terms on top of a learned model, to retain the ability to generate data without accumulating errors over long prediction horizons. We motivate this method theoretically and show that it counteracts an error term for model-based policy improvement. Experiments on MuJoCo- and PyBullet-benchmarks show that our method can drastically improve existing model-based approaches without introducing additional tuning parameters.

TAPEX: Table Pre-training via Learning a Neural SQL Executor

- Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, Jian-Guang Lou
- abstract@open-review(Poster): Recent progress in language model pre-training has achieved a great success via leveraging large-scale unstructured textual data. However, it is still a challenge to apply pre-training on structured tabular data due to the absence of large-scale high-quality tabular data. In this paper, we propose TAPEX to show that table pre-training can be achieved by learning a neural SQL executor over a synthetic corpus, which is obtained by automatically synthesizing executable SQL queries and their execution outputs. TAPEX addresses the data scarcity challenge via guiding the language model to mimic a SQL executor on the diverse, large-scale and high-quality synthetic corpus. We evaluate TAPEX on four benchmark datasets. Experimental results demonstrate that TAPEX outperforms previous table pre-training approaches by a large margin and achieves new state-of-the-art results on all of them. This includes the improvements on the weakly-supervised WikiSQL denotation accuracy to 89.5% (+2.3%), the WikiTableQuestions denotation accuracy to 57.5% (+4.8%), the SQA denotation accuracy to 74.5% (+3.5%), and the TabFact accuracy to 84.2% (+3.2%). To our knowledge, this is the first work to exploit table pre-training via synthetic executable programs and to achieve new state-of-the-art results on various downstream tasks. Our code can be found at https://github.com/microsoft/Table-Pretraining.

DARA: Dynamics-Aware Reward Augmentation in Offline Reinforcement Learning

- Jinxin Liu, Zhang Hongyin, Donglin Wang
- abstract@open-review(Poster): Offline reinforcement learning algorithms promise to be applicable in settings where a fixed dataset is available and no new experience can be acquired. However, such formulation is inevitably offline-data-hungry and, in practice, collecting a large offline dataset for one specific task over one specific environment is also costly and laborious. In this paper, we thus 1) formulate the offline dynamics adaptation by using (source) offline data collected from another dynamics to relax the requirement for the extensive (target) offline data, 2) characterize the dynamics shift problem in which prior offline methods do not scale well, and 3) derive a simple dynamics-aware reward augmentation (DARA) framework from both model-free and model-based offline settings. Specifically, DARA emphasizes learning from those source transition pairs that are adaptive for the target environment and mitigates the offline dynamics shift by characterizing state-action-next-state pairs instead of the typical state-action distribution sketched by prior offline RL methods. The experimental evaluation demonstrates that DARA, by augmenting rewards in the source offline dataset, can acquire an adaptive policy for the target environment and yet significantly reduce the requirement of target offline data. With only modest amounts of target offline data, our performance consistently outperforms the prior offline RL methods in both simulated and real-world tasks.

Explaining Point Processes by Learning Interpretable Temporal Logic Rules

- Shuang Li, Mingquan Feng, Lu Wang, Abdelmajid Essofi, Yufeng Cao, Junchi Yan, Le Song
- abstract@open-review(Poster): We propose a principled method to learn a set of human-readable logic rules to explain temporal point processes. We assume that the generative mechanisms underlying the temporal point processes are governed by a set of first-order temporal logic rules, as a compact representation of domain knowledge. Our method formulates the rule discovery process from noisy event data as a maximum likelihood problem, and designs an efficient and tractable branch-and-price algorithm to progressively search for new rules and expand existing rules. The proposed algorithm alternates between the rule generation stage and the rule evaluation stage, and uncovers the most important collection of logic rules within a fixed time limit for both synthetic and real event data. In a real healthcare application, we also had human experts (i.e., doctors) verify the learned temporal logic rules and provide further improvements. These expert-revised interpretable rules lead to a point process model which outperforms previous state-of-the-arts for symptom prediction, both in their occurrence times and types.

On Robust Prefix-Tuning for Text Classification

- Zonghan Yang, Yang Liu
- abstract@open-review(Poster): Recently, prefix-tuning has gained increasing attention as a parameter-efficient finetuning method for large-scale pretrained language models. The method keeps the pretrained models fixed and only updates the prefix token parameters for each downstream task. Despite being lightweight and modular, prefix-tuning still lacks robustness to textual adversarial attacks. However, most currently developed defense techniques necessitate auxiliary model update and storage, which inevitably hamper the modularity and low storage of prefix-tuning. In this work, we propose a robust prefix-tuning framework that preserves the efficiency and modularity of prefix-tuning. The core idea of our framework is leveraging the layerwise activations of the language model by correctly-classified training data as the standard for additional prefix finetuning. During the test phase, an extra batch-level prefix is tuned for each batch and added to the original prefix for robustness enhancement. Extensive experiments on three text classification benchmarks show that our framework substantially improves robustness over several strong baselines against five textual attacks of different types while maintaining comparable accuracy on clean texts. We also interpret our robust prefix-tuning framework from the optimal control perspective and pose several directions for future research.

Learning Graphon Mean Field Games and Approximate Nash Equilibria

- Kai Cui, Heinz Koeppl

- abstract@open-review(Poster): Recent advances at the intersection of dense large graph limits and mean field games have begun to enable the scalable analysis of a broad class of dynamical sequential games with large numbers of agents. So far, results have been largely limited to graphon mean field systems with continuous-time diffusive or jump dynamics, typically without control and with little focus on computational methods. We propose a novel discrete-time formulation for graphon mean field games as the limit of non-linear dense graph Markov games with weak interaction. On the theoretical side, we give extensive and rigorous existence and approximation properties of the graphon mean field solution in sufficiently large systems. On the practical side we provide general learning schemes for graphon mean field equilibria by either introducing agent equivalence classes or reformulating the graphon mean field system as a classical mean field system. By repeatedly finding a regularized optimal control solution and its generated mean field, we successfully obtain plausible approximate Nash equilibria in otherwise infeasible large dense graph games with many agents. Empirically, we are able to demonstrate on a number of examples that the finite-agent behavior comes increasingly close to the mean field behavior for our computed equilibria as the graph or system size grows, verifying our theory. More generally, we successfully apply policy gradient reinforcement learning in conjunction with sequential Monte Carlo methods.

Measuring CLEVRness: Black-box Testing of Visual Reasoning Models

- Spyridon Mouselinos, Henryk Michalewski, Mateusz Malinowski
- abstract@open-review(Poster): How can we measure the reasoning capabilities of intelligence systems? Visual question answering provides a convenient framework for testing the model's abilities by interrogating the model through questions about the scene. However, despite scores of various visual QA datasets and architectures, which sometimes yield even a super-human performance, the question of whether those architectures can actually reason remains open to debate. To answer this, we extend the visual question answering framework and propose the following behavioral test in the form of a two-player game. We consider black-box neural models of CLEVR. These models are trained on a diagnostic dataset benchmarking reasoning. Next, we train an adversarial player that re-configures the scene to fool the CLEVR model. We show that CLEVR models, which otherwise could perform at a ``human-level", can easily be fooled by our agent. Our results put in doubt whether data-driven approaches can do reasoning without exploiting the numerous biases that are often present in those datasets. Finally, we also propose a controlled experiment measuring the efficiency of such models to learn and perform reasoning.

Exploiting Class Activation Value for Partial-Label Learning

- Fei Zhang, Lei Feng, Bo Han, Tongliang Liu, Gang Niu, Tao Qin, Masashi Sugiyama
- abstract@open-review(Poster): Partial-label learning (PLL) solves the multi-class classification problem, where each training instance is assigned a set of candidate labels that include the true label. Recent advances showed that PLL can be compatible with deep neural networks, which achieved state-of-the-art performance. However, most of the existing deep PLL methods focus on designing proper training objectives under various assumptions on the collected data, which may limit their performance when the collected data cannot satisfy the adopted assumptions. In this paper, we propose to exploit the learned intrinsic representation of the model to identify the true label in the training process, which does not rely on any assumptions on the collected data. We make two key contributions. As the first contribution, we empirically show that the class activation map (CAM), a simple technique for discriminating the learning patterns of each class in images, could surprisingly be utilized to make accurate predictions on selecting the true label from candidate labels. Unfortunately, as CAM is confined to image inputs with convolutional neural networks, we are yet unable to directly leverage CAM to address the PLL problem with general inputs and models. Thus, as the second contribution, we propose the class activation value (CAV), which owns similar properties of CAM, while CAV is versatile in various types of inputs and models. Building upon CAV, we propose a novel method named CAV Learning (CAVL) that selects the true label by the class with the maximum CAV for model training. Extensive experiments on various datasets demonstrate that our proposed CAVL method achieves state-of-the-art performance.

Givens Coordinate Descent Methods for Rotation Matrix Learning in Trainable Embedding Indexes

- Yunjiang Jiang, Han Zhang, Yiming Qiu, Yun Xiao, Bo Long, Wen-Yun Yang
- abstract@open-review(Poster): Product quantization (PQ) coupled with a space rotation, is widely used in modern approximate nearest neighbor (ANN) search systems to significantly compress the disk storage for embeddings and speed up the inner product computation. Existing rotation learning methods, however, minimize quantization distortion for fixed embeddings, which are not applicable to an end-to-end training scenario where embeddings are updated constantly. In this paper, based on geometric intuitions from Lie group theory, in particular the special orthogonal groupSO(n), we propose a family of block Givens coordinate descent algorithms to learn rotation matrix that are provably convergent on any convex objectives. Compared to the state-of-the-art SVD method, the Givens algorithms are much more parallelizable, reducing runtime by orders of magnitude on modern GPUs, and converge more stably according to experimental studies. They further improve upon vanilla product quantization significantly in an end-to-end training scenario.

cosFormer: Rethinking Softmax In Attention

- Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, Yiran Zhong
- abstract@open-review(Poster): Transformer has shown great successes in natural language processing, computer vision, and audio processing. As one of its core components, the softmax attention helps to capture long-range dependencies yet prohibits its scale-up due to the quadratic space and time complexity to the sequence length. Kernel methods are often adopted to reduce the complexity by approximating the softmax operator. Nevertheless, due to the approximation errors, their performances vary in different tasks/corpus and suffer crucial performance drops when compared with the vanilla softmax attention. In this paper, we propose a linear transformer called cosFormer that can achieve comparable or better accuracy to the vanilla transformer in both casual and cross attentions. cosFormer is based on two key properties of softmax attention: i). non-negativeness of the attention matrix; ii). a non-linear re-weighting scheme that can concentrate the distribution of the attention matrix. As its linear substitute, cosFormer fulfills these properties with a linear operator and a cosine-based distance re-weighting mechanism. Extensive experiments on language modeling and text understanding tasks demonstrate the effectiveness of our method. We further examine our method on long sequences and achieve state-of-the-art performance on the Long-Range Arena benchmark. The source code is available at https://github.com/OpenNLPLab/cosFormer.

FALCON: Fast Visual Concept Learning by Integrating Images, Linguistic descriptions, and Conceptual Relations

- Lingjie Mei, Jiayuan Mao, Ziqi Wang, Chuang Gan, Joshua B. Tenenbaum
- abstract@open-review(Poster): We present a meta-learning framework for learning new visual concepts quickly, from just one or a few examples, guided by multiple naturally occurring data streams: simultaneously looking at images, reading sentences that describe the objects in the scene, and interpreting supplemental sentences that relate the novel concept with other concepts. The learned concepts support downstream applications, such as answering questions by reasoning about unseen images. Our model, namely FALCON, represents individual visual concepts, such as colors and shapes, as axis-aligned boxes in a high-dimensional space (the `box embedding space''). Given an input image and its paired sentence, our model first resolves the referential expression in the sentence and associates the novel concept with particular objects in the scene. Next, our model interprets supplemental sentences to relate the novel concept with other known concepts, such as`X has property Y" or ``X is a kind of Y". Finally, it infers an optimal box embedding for the novel concept that jointly 1) maximizes the likelihood of the observed instances in the image, and 2) satisfies the relationships between the novel concepts and the known ones. We demonstrate the effectiveness of our model on both synthetic and real-world datasets.

HyAR: Addressing Discrete-Continuous Action Reinforcement Learning via Hybrid Action Representation

- Boyan Li, Hongyao Tang, YAN ZHENG, Jianye HAO, Pengyi Li, Zhen Wang, Zhaopeng Meng, LI Wang
- abstract@open-review(Poster): Discrete-continuous hybrid action space is a natural setting in many practical problems, such as robot control and game AI. However, most previous Reinforcement Learning (RL) works only demonstrate the success in controlling with either discrete or continuous action space, while seldom take into account the hybrid action space. One naive way to address hybrid action RL is to convert the hybrid action space into a unified homogeneous action space by discretization or continualization, so that conventional RL algorithms can be applied. However, this ignores the underlying structure of hybrid action space and also induces the scalability issue and additional approximation difficulties, thus leading to degenerated results. In this paper, we propose Hybrid Action Representation (HyAR) to learn a compact and decodable latent representation space for the original hybrid action space. HyAR constructs the latent space and embeds the dependence between discrete action and continuous parameter via an embedding table and conditional Variantional Auto-Encoder (VAE). To further improve the effectiveness, the action representation is trained to be semantically smooth through unsupervised environmental dynamics prediction. Finally, the agent then learns its policy with conventional DRL algorithms in the learned representation space and interacts with the environment by decoding the hybrid action embeddings to the original action space. We evaluate HyAR in a variety of environments with discrete-continuous action space. The results demonstrate the superiority of HyAR when compared with previous baselines, especially for high-dimensional action spaces.

Transferable Adversarial Attack based on Integrated Gradients

- Yi Huang, Adams Wai-Kin Kong
- abstract@open-review(Poster): The vulnerability of deep neural networks to adversarial examples has drawn tremendous attention from the community. Three approaches, optimizing standard objective functions, exploiting attention maps, and smoothing decision surfaces, are commonly used to craft adversarial examples. By tightly integrating the three approaches, we propose a new and simple algorithm named Transferable Attack based on Integrated Gradients (TAIG) in this paper, which can find highly transferable adversarial examples for black-box attacks. Unlike previous methods using multiple computational terms or combining with other methods, TAIG integrates the three approaches into one single term. Two versions of TAIG that compute their integrated gradients on a straight-line path and a random piecewise linear path are studied. Both versions offer strong transferability and can seamlessly work together with the previous methods. Experimental results demonstrate that TAIG outperforms the state-of-the-art methods.

How to deal with missing data in supervised deep learning?

- Niels Bruun Ipsen, Pierre-Alexandre Mattei, Jes Frellsen
- abstract@open-review(Poster): The issue of missing data in supervised learning has been largely overlooked, especially in the deep learning community. We investigate strategies to adapt neural architectures for handling missing values. Here, we focus on regression and classification problems where the features are assumed to be missing at random. Of particular interest are schemes that allow reusing as-is a neural discriminative architecture. To address supervised deep learning with missing values, we propose to marginalize over missing values in a joint model of covariates and outcomes. Thereby, we leverage both the flexibility of deep generative models to describe the distribution of the covariates and the power of purely discriminative models to make predictions. More precisely, a deep latent variable model can be learned jointly with the discriminative model, using importance-weighted variational inference, essentially using importance sampling to mimick averaging over multiple imputations. In low-capacity regimes, or when the discriminative model has a strong inductive bias, we find that our hybrid generative/discriminative approach generally outperforms single imputations methods.

Topological Graph Neural Networks

- Max Horn, Edward De Brouwer, Michael Moor, Yves Moreau, Bastian Rieck, Karsten Borgwardt
- abstract@open-review(Poster): Graph neural networks (GNNs) are a powerful architecture for tackling graph learning tasks, yet have been shown to be oblivious to eminent substructures such as cycles. We present TOGL, a novel layer that incorporates global topological information of a graph using persistent homology. TOGL can be easily integrated into any type of GNN and is strictly more expressive (in terms the Weisfeilerâ€“Lehman graph isomorphism test) than message-passing GNNs. Augmenting GNNs with TOGL leads to improved predictive performance for graph and node classification tasks, both on synthetic data sets, which can be classified by humans using their topology but not by ordinary GNNs, and on real-world data.

Learning Value Functions from Undirected State-only Experience

- Matthew Chang, Arjun Gupta, Saurabh Gupta
- abstract@open-review(Poster): This paper tackles the problem of learning value functions from undirected state-only experience (state transitions without action labels i.e. (s,s',r) tuples). We first theoretically characterize the applicability of Q-learning in this setting. We show that tabular Q-learning in discrete Markov decision processes (MDPs) learns the same value function under any arbitrary refinement of the action space. This theoretical result motivates the design of Latent Action Q-learning or LAQ, an offline RL method that can learn effective value functions from state-only experience. Latent Action Q-learning (LAQ) learns value functions using Q-learning on discrete latent actions obtained through a latent-variable future prediction model. We show that LAQ can recover value functions that have high correlation with value functions learned using ground truth actions. Value functions learned using LAQ lead to sample efficient acquisition of goal-directed behavior, can be used with domain-specific low-level controllers, and facilitate transfer across embodiments. Our experiments in 5 environments ranging from 2D grid world to 3D visual navigation in realistic environments demonstrate the benefits of LAQ over simpler alternatives, imitation learning oracles, and competing methods.

The Boltzmann Policy Distribution: Accounting for Systematic Suboptimality in Human Models

- Cassidy Laidlaw, Anca Dragan
- abstract@open-review(Poster): Models of human behavior for prediction and collaboration tend to fall into two categories: ones that learn from large amounts of data via imitation learning, and ones that assume human behavior to be noisily-optimal for some reward function. The former are very useful, but only when it is possible to gather a lot of human data in the target environment and distribution. The advantage of the latter type, which includes Boltzmann rationality, is the ability to make accurate predictions in new environments without extensive data when humans are actually close to optimal. However, these models fail when humans exhibit systematic suboptimality, i.e. when their deviations from optimal behavior are not independent, but instead consistent over time. Our key insight is that systematic suboptimality can be modeled by predicting policies, which couple action choices over time, instead of trajectories. We introduce the Boltzmann policy distribution (BPD), which serves as a prior over human policies and adapts via Bayesian inference to capture systematic deviations by observing human actions during a single episode. The BPD is difficult to compute and represent because policies lie in a high-dimensional continuous space, but we leverage tools from generative and sequence modeling to enable efficient sampling and inference. We show that the BPD enables prediction of human behavior and human-AI collaboration equally as well as imitation learning-based human models while using far less data.

WeakM3D: Towards Weakly Supervised Monocular 3D Object Detection

- Liang Peng, Senbo Yan, Boxi Wu, Zheng Yang, Xiaofei He, Deng Cai
- abstract@open-review(Poster): Monocular 3D object detection is one of the most challenging tasks in 3D scene understanding. Due to the ill-posed nature of monocular imagery, existing monocular 3D detection methods highly rely on training with the manually annotated 3D box labels on the LiDAR point clouds. This annotation process is very laborious and expensive. To dispense with the reliance on 3D box labels, in this paper we explore the weakly supervised monocular 3D detection. Specifically, we first detect 2D boxes on the image. Then, we adopt the generated 2D boxes to select corresponding RoI LiDAR points as the weak supervision. Eventually, we adopt a network to predict 3D boxes which can tightly align with associated RoI LiDAR

points. This network is learned by minimizing our newly-proposed 3D alignment loss between the 3D box estimates and the corresponding RoI LiDAR points. We will illustrate the potential challenges of the above learning problem and resolve these challenges by introducing several effective designs into our method. Codes are available at https://github.com/SPengLiang/WeakM3D.

Exploring Memorization in Adversarial Training

- Yinpeng Dong, Ke Xu, Xiao Yang, Tianyu Pang, Zhijie Deng, Hang Su, Jun Zhu
- abstract@open-review(Poster): Deep learning models have a propensity for fitting the entire training set even with random labels, which requires memorization of every training sample. In this paper, we explore the memorization effect in adversarial training (AT) for promoting a deeper understanding of model capacity, convergence, generalization, and especially robust overfitting of the adversarially trained models. We first demonstrate that deep networks have sufficient capacity to memorize adversarial examples of training data with completely random labels, but not all AT algorithms can converge under the extreme circumstance. Our study of AT with random labels motivates further analyses on the convergence and generalization of AT. We find that some AT approaches suffer from a gradient instability issue and the recently suggested complexity measures cannot explain robust generalization by considering models trained on random labels. Furthermore, we identify a significant drawback of memorization in AT that it could result in robust overfitting. We then propose a new mitigation algorithm motivated by detailed memorization analyses. Extensive experiments on various datasets validate the effectiveness of the proposed method.

Disentanglement Analysis with Partial Information Decomposition

- Seiya Tokui, Issei Sato
- abstract@open-review(Poster): We propose a framework to analyze how multivariate representations disentangle ground-truth generative factors. A quantitative analysis of disentanglement has been based on metrics designed to compare how one variable explains each generative factor. Current metrics, however, may fail to detect entanglement that involves more than two variables, e.g., representations that duplicate and rotate generative factors in high dimensional spaces. In this work, we establish a framework to analyze information sharing in a multivariate representation with Partial Information Decomposition and propose a new disentanglement metric. This framework enables us to understand disentanglement in terms of uniqueness, redundancy, and synergy. We develop an experimental protocol to assess how increasingly entangled representations are evaluated with each metric and confirm that the proposed metric correctly responds to entanglement. Through experiments on variational autoencoders, we find that models with similar disentanglement scores have a variety of characteristics in entanglement, for each of which a distinct strategy may be required to obtain a disentangled representation.

Differentiable Gradient Sampling for Learning Implicit 3D Scene Reconstructions from a Single Image

- Shizhan Zhu, Sayna Ebrahimi, Angjoo Kanazawa, Trevor Darrell
- abstract@open-review(Poster): Implicit shape models are promising 3D representations for modeling arbitrary locations, with Signed Distance Functions (SDFs) particularly suitable for clear mesh surface reconstruction. Existing approaches for single object reconstruction impose supervision signals based on the loss of the signed distance value from all locations in a scene, posing difficulties when extending to real-world scenarios. The spatial gradient of the signed distance field, rather than the SDF value itself, has not been typically employed as a source of supervision for single-view reconstruction, in part due to the difficulties of differentiable sampling a spatial gradient from the feature map. In this study, we derive a novel closed-form gradient sampling solution for Differentialble Gradient Sampling (DGS) that enables backpropagation of the loss of the spatial gradient back to the feature map pixels, thus allowing the imposition of the loss efficiently on the spatial gradient. As a result, we achieve high-quality single view indoor scene reconstruction results learning directly from a real-world scanned dataset (e.g. ScannetV2). Our model also performs well when generalizing to unseen images downloaded directly from the internet (Fig. 1). We comfortably advanced the state-of-the-art results with several established datasets including ShapeNet and ScannetV2; extensive quantitative analysis confirmed that our proposed DGS module plays an essential role in achieving this performance improvement. Full codes are available in MaskedURL.

Learning Continuous Environment Fields via Implicit Functions

- Xueting Li, Shalini De Mello, Xiaolong Wang, Ming-Hsuan Yang, Jan Kautz, Sifei Liu
- abstract@open-review(Poster): We propose a novel scene representation that encodes reaching distance -- the distance between any position in the scene to a goal along a feasible trajectory. We demonstrate that this environment field representation can directly guide the dynamic behaviors of agents in 2D mazes or 3D indoor scenes. Our environment field is a continuous representation and learned via a neural implicit function using discretely sampled training data. We showcase its application for agent navigation in 2D mazes, and human trajectory prediction in 3D indoor environments. To produce physically plausible and natural trajectories for humans, we additionally learn a generative model that predicts regions where humans commonly appear, and enforce the environment field to be defined within such regions. Extensive experiments demonstrate that the proposed method can generate both feasible and plausible trajectories efficiently and accurately.

Causal Contextual Bandits with Targeted Interventions

- Chandrasekar Subramanian, Balaraman Ravindran
- abstract@open-review(Poster): We study a contextual bandit setting where the learning agent has the ability to perform interventions on targeted subsets of the population, apart from possessing qualitative causal side-information. This novel formalism captures intricacies in real-world scenarios such as software product experimentation where targeted experiments can be conducted. However, this fundamentally changes the set of options that the agent has, compared to standard contextual bandit settings, necessitating new techniques. This is also the first work that integrates causal side-information in a contextual bandit setting, where the agent aims to learn a policy that maps contexts to arms (as opposed to just identifying one best arm). We propose a new algorithm, which we show empirically performs better than baselines on experiments that use purely synthetic data and on real world-inspired experiments. We also prove a bound on regret that theoretically guards performance.

Sound and Complete Neural Network Repair with Minimality and Locality Guarantees

- Feisi Fu, Wenchao Li
- abstract@open-review(Poster): We present a novel methodology for repairing neural networks that use ReLU activation functions. Unlike existing methods that rely on modifying the weights of a neural network which can induce a global change in the function space, our approach applies only a localized change in the function space while still guaranteeing the removal of the buggy behavior. By leveraging the piecewise linear nature of ReLU networks, our approach can efficiently construct a patch network tailored to the linear region where the buggy input resides, which when combined with the original network, provably corrects the behavior on the buggy input. Our method is both sound and complete -- the repaired network is guaranteed to fix the buggy input, and a patch is guaranteed to be found for any buggy input. Moreover, our approach preserves the continuous piecewise linear nature of ReLU networks, automatically generalizes the repair to all the points including other undetected buggy inputs inside the repair region, is minimal in terms of changes in the function space, and guarantees that outputs on inputs away from the repair region are unaltered. On several benchmarks, we show that our approach significantly outperforms existing methods in terms of locality and limiting negative side effects.

Blaschke Product Neural Networks (BPNN): A Physics-Infused Neural Network for Phase Retrieval of Meromorphic Functions

- Juncheng Dong, Simiao Ren, Yang Deng, Omar Khatib, Jordan Malof, Mohammadreza Soltani, Willie Padilla, Vahid Tarokh
- abstract@open-review(Poster): Numerous physical systems are described by ordinary or partial differential equations whose solutions are given by holomorphic or meromorphic functions in the complex domain. In many cases, only the magnitude of these functions are observed on various points on the purely imaginary $j\omega$-axis since coherent measurement of their phases is often expensive. However, it is desirable to retrieve the lost phases from the magnitudes when possible. To this end, we propose a physics-infused deep neural network based on the Blaschke products for phase retrieval. Inspired by the Helson and Sarason Theorem, we recover coefficients of a rational function of Blaschke products using a Blaschke Product Neural Network (BPNN), based upon the magnitude observations as input. The resulting rational function is then used for phase retrieval. We compare the BPNN to conventional deep neural networks (NNs) on several phase retrieval problems, comprising both synthetic and contemporary real-world problems (e.g., metamaterials for which data collection requires substantial expertise and is time consuming). On each phase retrieval problem, we compare against a population of conventional NNs of varying size and hyperparameter settings. Even without any hyper-parameter search, we find that BPNNs consistently outperform the population of optimized NNs in scarce data scenarios, and do so despite being much smaller models. The results can in turn be applied to calculate the refractive index of metamaterials, which is an important problem in emerging areas of material science.

Automated Self-Supervised Learning for Graphs

- Wei Jin, Xiaorui Liu, Xiangyu Zhao, Yao Ma, Neil Shah, Jiliang Tang
- abstract@open-review(Poster): Graph self-supervised learning has gained increasing attention due to its capacity to learn expressive node representations. Many pretext tasks, or loss functions have been designed from distinct perspectives. However, we observe that different pretext tasks affect downstream tasks differently cross datasets, which suggests that searching pretext tasks is crucial for graph self-supervised learning. Different from existing works focusing on designing single pretext tasks, this work aims to investigate how to automatically leverage multiple pretext tasks effectively. Nevertheless, evaluating representations derived from multiple pretext tasks without direct access to ground truth labels makes this problem challenging. To address this obstacle, we make use of a key principle of many real-world graphs, i.e., homophily, or the principle that ``like attracts like,'' as the guidance to effectively search various self-supervised pretext tasks. We provide theoretical understanding and empirical evidence to justify the flexibility of homophily in this search task. Then we propose the AutoSSL framework which can automatically search over combinations of various self-supervised tasks. By evaluating the framework on 7 real-world datasets, our experimental results show that AutoSSL can significantly boost the performance on downstream tasks including node clustering and node classification compared with training under individual tasks.

Creating Training Sets via Weak Indirect Supervision

- Jieyu Zhang, Bohan Wang, Xiangchen Song, Yujing Wang, Yaming Yang, Jing Bai, Alexander Ratner
- abstract@open-review(Poster): Creating labeled training sets has become one of the major roadblocks in machine learning. To address this, recent Weak Supervision (WS) frameworks synthesize training labels from multiple potentially noisy supervision sources. However, existing frameworks are restricted to supervision sources that share the same output space as the target task. To extend the scope of usable sources, we formulate Weak Indirect Supervision (WIS), a new research problem for automatically synthesizing training labels based on indirect supervision sources that have different output label spaces. To overcome the challenge of mismatched output spaces, we develop a probabilistic modeling approach, PLRM, which uses user-provided label relations to model and leverage indirect supervision sources. Moreover, we provide a theoretically-principled test of the distinguishability of PLRM for unseen labels, along with an generalization bound. On both image and text classification tasks as well as an industrial advertising application, we demonstrate the advantages of PLRM by outperforming baselines by a margin of 2%-9%.

Do Not Escape From the Manifold: Discovering the Local Coordinates on the Latent Space of GANs

- Jaewoong Choi, Junho Lee, Changyeon Yoon, Jung Ho Park, Geonho Hwang, Myungjoo Kang
- abstract@open-review(Poster): The discovery of the disentanglement properties of the latent space in GANs motivated a lot of research to find the semantically meaningful directions on it. In this paper, we suggest that the disentanglement property is closely related to the geometry of the latent space. In this regard, we propose an unsupervised method for finding the semantic-factorizing directions on the intermediate latent space of GANs based on the local geometry. Intuitively, our proposed method, called $\textit{Local Basis}$, finds the principal variation of the latent space in the neighborhood of the base latent variable. Experimental results show that the local principal variation corresponds to the semantic factorization and traversing along it provides strong robustness to image traversal. Moreover, we suggest an explanation for the limited success in finding the global traversal directions in the latent space, especially $\mathcal{W}$-space of StyleGAN2. We show that $\mathcal{W}$-space is warped globally by comparing the local geometry, discovered from Local Basis, through the metric on Grassmannian Manifold. The global warpage implies that the latent space is not well-aligned globally and therefore the global traversal directions are bound to show limited success on it.

GradSign: Model Performance Inference with Theoretical Insights

- Zhihao Zhang, Zhihao Jia
- abstract@open-review(Poster): A key challenge in neural architecture search (NAS) is quickly inferring the predictive performance of a broad spectrum of networks to discover statistically accurate and computationally efficient ones. We refer to this task as model performance inference (MPI). The current practice for efficient MPI is gradient-based methods that leverage the gradients of a network at initialization to infer its performance. However, existing gradient-based methods rely only on heuristic metrics and lack the necessary theoretical foundations to consolidate their designs. We propose GradSign, an accurate, simple, and flexible metric for model performance inference with theoretical insights. The key idea behind GradSign is a quantity Î¨ to analyze the sample-wise optimization landscape of different networks. Theoretically, we show that Î¨ is an upper bound for both the training and true population losses of a neural network under reasonable assumptions. However, it is computationally prohibitive to directly calculate Î¨ for modern neural networks. To address this challenge, we design GradSign, an accurate and simple approximation of Î¨ using the gradients of a network evaluated at a random initialization state. Evaluation on seven NAS benchmarks across three training datasets shows that GradSign generalizes well to real-world networks and consistently outperforms state-of-the-art gradient-based methods for MPI evaluated by Spearmanâ€™s Ï and Kendallâ€™s Tau. Additionally, we integrate GradSign into four existing NAS algorithms and show that the GradSign-assisted NAS algorithms outperform their vanilla counterparts by improving the accuracies of best-discovered networks by up to 0.3%, 1.1%, and 1.0% on three real-world tasks. Code is available at https://github.com/JackFram/GradSign

You are AllSet: A Multiset Function Framework for Hypergraph Neural Networks

- Eli Chien, Chao Pan, Jianhao Peng, Olgica Milenkovic
- abstract@open-review(Poster): Hypergraphs are used to model higher-order interactions amongst agents and there exist many practically relevant instances of hypergraph datasets. To enable the efficient processing of hypergraph data, several hypergraph neural network platforms have been proposed for learning hypergraph properties and structure, with a special focus on node classification tasks. However, almost all existing methods use heuristic propagation rules and offer suboptimal performance on benchmarking datasets. We propose AllSet, a new hypergraph neural network paradigm that represents a highly general framework for (hyper)graph neural networks and for the first time implements hypergraph neural network layers as compositions of two multiset functions that can be efficiently learned for each task and each dataset. The proposed AllSet framework also for the first time integrates Deep Sets and Set Transformers with hypergraph neural networks for the purpose of learning multiset functions and therefore allows for significant modeling flexibility and high expressive power. To evaluate the performance of AllSet, we conduct the most extensive experiments to date involving ten known benchmarking datasets and three newly curated datasets that represent significant challenges for hypergraph node classification. The results demonstrate that our method has the unique ability to either match or outperform all other hypergraph neural networks across the tested datasets:

As an example, the performance improvements over existing methods and a new method based on heterogeneous graph neural networks are close to 4% on the Yelp and Zoo datasets, and 3% on the Walmart dataset.

Synchromesh: Reliable Code Generation from Pre-trained Language Models

- Gabriel Poesia, Alex Polozov, Vu Le, Ashish Tiwari, Gustavo Soares, Christopher Meek, Sumit Gulwani
- abstract@open-review(Poster): Large pre-trained language models have been used to generate code, providing a flexible interface for synthesizing programs from natural language specifications. However, they often violate syntactic and semantic rules of their output language, limiting their practical usability. In this paper, we propose Synchromesh: a framework for substantially improving the reliability of pre-trained models for code generation. Synchromesh comprises two components. First, it retrieves few-shot examples from a training bank using Target Similarity Tuning (TST), a novel method for semantic example selection. TST learns to recognize utterances that describe similar target programs despite of differences in surface natural language features. Then, Synchromesh feeds the examples to a pre-trained language model and samples programs using Constrained Semantic Decoding (CSD): a general framework for constraining the output to a set of valid programs in the target language. CSD leverages constraints on partial outputs to sample complete correct programs, and needs neither re-training nor fine-tuning of the language model. We evaluate our methods by synthesizing code from natural language descriptions using GPT-3 and Codex in three real-world languages: SQL queries, Vega-Lite visualizations and SMCalFlow programs. These domains showcase rich constraints that CSD is able to enforce, including syntax, scoping and typing rules. Across all languages, we observe complementary gains from CSD and TST in prediction accuracy and in effectively preventing parsing, type and run-time errors.

Learning curves for continual learning in neural networks: Self-knowledge transfer and forgetting

- Ryo Karakida, Shotaro Akaho
- abstract@open-review(Poster): Sequential training from task to task is becoming one of the major objects in deep learning applications such as continual learning and transfer learning. Nevertheless, it remains unclear under what conditions the trained model's performance improves or deteriorates. To deepen our understanding of sequential training, this study provides a theoretical analysis of generalization performance in a solvable case of continual learning. We consider neural networks in the neural tangent kernel (NTK) regime that continually learn target functions from task to task, and investigate the generalization by using an established statistical mechanical analysis of kernel ridge-less regression. We first show characteristic transitions from positive to negative transfer. More similar targets above a specific critical value can achieve positive knowledge transfer for the subsequent task while catastrophic forgetting occurs even with very similar targets. Next, we investigate a variant of continual learning which supposes the same target function in multiple tasks. Even for the same target, the trained model shows some transfer and forgetting depending on the sample size of each task. We can guarantee that the generalization error monotonically decreases from task to task for equal sample sizes while unbalanced sample sizes deteriorate the generalization. We respectively refer to these improvement and deterioration as self-knowledge transfer and forgetting, and empirically confirm them in realistic training of deep neural networks as well.

Energy-Based Learning for Cooperative Games, with Applications to Valuation Problems in Machine Learning

- Yatao Bian, Yu Rong, Tingyang Xu, Jiaxiang Wu, Andreas Krause, Junzhou Huang
- abstract@open-review(Poster): Valuation problems, such as feature interpretation, data valuation and model valuation for ensembles, become increasingly more important in many machine learning applications. Such problems are commonly solved by well-known game-theoretic criteria, such as Shapley value or Banzhaf value. In this work, we present a novel energy-based treatment for cooperative games, with a theoretical justification by the maximum entropy framework. Surprisingly, by conducting variational inference of the energy-based model, we recover various game-theoretic valuation criteria through conducting one-step fixed point iteration for maximizing the mean-field ELBO objective. This observation also verifies the rationality of existing criteria, as they are all attempting to decouple the correlations among the players through the mean-field approach. By running fixed point iteration for multiple steps, we achieve a trajectory of the valuations, among which we define the valuation with the best conceivable decoupling error as the Variational Index. We prove that under uniform initializations, these variational valuations all satisfy a set of game-theoretic axioms. We experimentally demonstrate that the proposed Variational Index enjoys lower decoupling error and better valuation performance on certain synthetic and real-world valuation problems.

Pessimistic Model-based Offline Reinforcement Learning under Partial Coverage

- Masatoshi Uehara, Wen Sun
- abstract@open-review(Poster): We study model-based offline Reinforcement Learning with general function approximation without a full coverage assumption on the offline data distribution. We present an algorithm named Constrained Pessimistic Policy Optimization (CPPO) which leverages a general function class and uses a constraint over the models to encode pessimism. Under the assumption that the ground truth model belongs to our function class (i.e., realizability in the function class), CPPO has a PAC guarantee with offline data only providing partial coverage, i.e., it can learn a policy that competes against any policy covered by the offline data. We then demonstrate that this algorithmic framework can be applied to many specialized Markov Decision Processes where the additional structural assumptions can further refine the concept of partial coverage. Two notable examples are: (1) low- rank MDP with representation learning where the partial coverage condition is defined using a relative condition number measured by the unknown ground truth feature representation; (2) factored MDP where the partial coverage condition is defined using density-ratio based concentrability coefficients associated with individual factors.

Cold Brew: Distilling Graph Node Representations with Incomplete or Missing Neighborhoods

- Wenqing Zheng, Edward W Huang, Nikhil Rao, Sumeet Katariya, Zhangyang Wang, Karthik Subbian
- abstract@open-review(Poster): Graph Neural Networks (GNNs) have achieved state-of-the-art performance in node classification, regression, and recommendation tasks. GNNs work well when rich and high-quality connections are available. However, their effectiveness is often jeopardized in many real-world graphs in which node degrees have power-law distributions. The extreme case of this situation, where a node may have no neighbors, is called Strict Cold Start (SCS). SCS forces the prediction to rely completely on the node's own features. We propose Cold Brew, a teacher-student distillation approach to address the SCS and noisy-neighbor challenges for GNNs. We also introduce feature contribution ratio (FCR), a metric to quantify the behavior of inductive GNNs to solve SCS. We experimentally show that FCR disentangles the contributions of different graph data components and helps select the best architecture for SCS generalization. We further demonstrate the superior performance of Cold Brew on several public benchmark and proprietary e-commerce datasets, where many nodes have either very few or noisy connections. Our source code is available at https://github.com/amazon-research/gnn-tail-generalization.

NASI: Label- and Data-agnostic Neural Architecture Search at Initialization

- Yao Shu, Shaofeng Cai, Zhongxiang Dai, Beng Chin Ooi, Bryan Kian Hsiang Low
- abstract@open-review(Poster): Recent years have witnessed a surging interest in Neural Architecture Search (NAS). Various algorithms have been proposed to improve the search efficiency and effectiveness of NAS, i.e., to reduce the search cost and improve the generalization performance of the selected architectures, respectively. However, the search efficiency of these algorithms is severely limited by the need for model training during the search process. To overcome this limitation, we propose a novel NAS algorithm called NAS at Initialization (NASI) that exploits the capability of a Neural Tangent Kernel in being able to characterize the performance of candidate architectures at initialization, hence allowing model training to be completely avoided to boost the search efficiency. Besides the improved search efficiency, NASI also achieves competitive search effectiveness on various datasets

like CIFAR-10/100 and ImageNet. Further, NASI is shown to be label- and data-agnostic under mild conditions, which guarantees the transferability of architectures selected by our NASI over different datasets.

How to Train Your MAML to Excel in Few-Shot Classification

- Han-Jia Ye, Wei-Lun Chao
- abstract@open-review(Poster): Model-agnostic meta-learning (MAML) is arguably one of the most popular meta-learning algorithms nowadays. Nevertheless, its performance on few-shot classification is far behind many recent algorithms dedicated to the problem. In this paper, we point out several key facets of how to train MAML to excel in few-shot classification. First, we find that MAML needs a large number of gradient steps in its inner loop update, which contradicts its common usage in few-shot classification. Second, we find that MAML is sensitive to the class label assignments during meta-testing. Concretely, MAML meta-trains the initialization of an N-way classifier. These N ways, during meta-testing, then have "$N!$" different permutations to be paired with a few-shot task of N novel classes. We find that these permutations lead to a huge variance of accuracy, making MAML unstable in few-shot classification. Third, we investigate several approaches to make MAML permutation-invariant, among which meta-training a single vector to initialize all the N weight vectors in the classification head performs the best. On benchmark datasets like MiniImageNet and TieredImageNet, our approach, which we name UNICORN-MAML, performs on a par with or even outperforms many recent few-shot classification algorithms, without sacrificing MAML's simplicity.

Communication-Efficient Actor-Critic Methods for Homogeneous Markov Games

- Dingyang Chen, Yile Li, Qi Zhang
- abstract@open-review(Poster): Recent success in cooperative multi-agent reinforcement learning (MARL) relies on centralized training and policy sharing. Centralized training eliminates the issue of non-stationarity MARL yet induces large communication costs, and policy sharing is empirically crucial to efficient learning in certain tasks yet lacks theoretical justification. In this paper, we formally characterize a subclass of cooperative Markov games where agents exhibit a certain form of homogeneity such that policy sharing provably incurs no suboptimality. This enables us to develop the first consensus-based decentralized actor-critic method where the consensus update is applied to both the actors and the critics while ensuring convergence. We also develop practical algorithms based on our decentralized actor-critic method to reduce the communication cost during training, while still yielding policies comparable with centralized training.

MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer

- Sachin Mehta, Mohammad Rastegari
- abstract@open-review(Poster): Light-weight convolutional neural networks (CNNs) are the de-facto for mobile vision tasks. Their spatial inductive biases allow them to learn representations with fewer parameters across different vision tasks. However, these networks are spatially local. To learn global representations, self-attention-based vision trans-formers (ViTs) have been adopted. Unlike CNNs, ViTs are heavy-weight. In this paper, we ask the following question: is it possible to combine the strengths of CNNs and ViTs to build a light-weight and low latency network for mobile vision tasks? Towards this end, we introduce MobileViT, a light-weight and general-purpose vision transformer for mobile devices. MobileViT presents a different perspective for the global processing of information with transformers, i.e., transformers as convolutions. Our results show that MobileViT significantly outperforms CNN- and ViT-based networks across different tasks and datasets. On the ImageNet-1k dataset, MobileViT achieves top-1 accuracy of 78.4% with about 6 million parameters, which is 3.2% and 6.2% more accurate than MobileNetv3 (CNN-based) and DeIT (ViT-based) for a similar number of parameters. On the MS-COCO object detection task, MobileViT is 5.7% more accurate than MobileNetv3 for a similar number of parameters.

Our source code is open-source and available at: https://github.com/apple/ml-cvnets

Spatial Graph Attention and Curiosity-driven Policy for Antiviral Drug Discovery

- Yulun Wu, Nicholas Choma, Andrew Deru Chen, Mikaela Cashman, Erica Teixeira Prates, Veronica G Melesse Vergara, Manesh B Shah, Austin Clyde, Thomas Brettin, Wibe Albert de Jong, Neeraj Kumar, Martha S Head, Rick L. Stevens, Peter Nugent, Daniel A Jacobson, James B Brown
- abstract@open-review(Poster): We developed Distilled Graph Attention Policy Network (DGAPN), a reinforcement learning model to generate novel graph-structured chemical representations that optimize user-defined objectives by efficiently navigating a physically constrained domain. The framework is examined on the task of generating molecules that are designed to bind, noncovalently, to functional sites of SARS-CoV-2 proteins. We present a spatial Graph Attention (sGAT) mechanism that leverages self-attention over both node and edge attributes as well as encoding the spatial structure --- this capability is of considerable interest in synthetic biology and drug discovery. An attentional policy network is introduced to learn the decision rules for a dynamic, fragment-based chemical environment, and state-of-the-art policy gradient techniques are employed to train the network with stability. Exploration is driven by the stochasticity of the action space design and the innovation reward bonuses learned and proposed by random network distillation. In experiments, our framework achieved outstanding results compared to state-of-the-art algorithms, while reducing the complexity of paths to chemical synthesis.

Surrogate NAS Benchmarks: Going Beyond the Limited Search Spaces of Tabular NAS Benchmarks

- Arber Zela, Julien Niklas Siems, Lucas Zimmer, Jovita Lukasik, Margret Keuper, Frank Hutter
- abstract@open-review(Poster): The most significant barrier to the advancement of Neural Architecture Search (NAS) is its demand for large computational resources, which hinders scientifically sound empirical evaluations of NAS methods. Tabular NAS benchmarks have alleviated this problem substantially, making it possible to properly evaluate NAS methods in seconds on commodity machines. However, an unintended consequence of tabular NAS benchmarks has been a focus on extremely small architectural search spaces since their construction relies on exhaustive evaluations of the space. This leads to unrealistic results that do not transfer to larger spaces. To overcome this fundamental limitation, we propose a methodology to create cheap NAS surrogate benchmarks for arbitrary search spaces. We exemplify this approach by creating surrogate NAS benchmarks on the existing tabular NAS-Bench-101 and on two widely used NAS search spaces with up to 10^{21} architectures (10^{13} times larger than any previous tabular NAS benchmark). We show that surrogate NAS benchmarks can model the true performance of architectures better than tabular benchmarks (at a small fraction of the cost), that they lead to faithful estimates of how well different NAS methods work on the original non-surrogate benchmark, and that they can generate new scientific insight. We open-source all our code and believe that surrogate NAS benchmarks are an indispensable tool to extend scientifically sound work on NAS to large and exciting search spaces.

Certified Robustness for Deep Equilibrium Models via Interval Bound Propagation

- Colin Wei, J Zico Kolter
- abstract@open-review(Poster): Deep equilibrium layers (DEQs) have demonstrated promising performance and are competitive with standard explicit models on many benchmarks. However, little is known about certifying robustness for these models. Inspired by interval bound propagation (IBP), we propose the IBP-MonDEQ layer, a DEQ layer whose robustness can be verified by computing upper and lower interval bounds on the output. Our key insights are that these interval bounds can be obtained as the fixed-point solution to an IBP-inspired equilibrium equation, and furthermore, that this solution always exists and is unique when the layer obeys a certain parameterization. This fixed point can be interpreted as the result of applying IBP to an infinitely deep, weight-tied neural network, which may be of independent interest, as IBP bounds are typically unstable for deeper networks. Our

empirical comparison reveals that models with IBP-MonDEQ layers can achieve comparable ℓ_{∞} certified robustness to similarly-sized fully explicit networks.

Crystal Diffusion Variational Autoencoder for Periodic Material Generation

- Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, Tommi S. Jaakkola
- abstract@open-review(Poster): Generating the periodic structure of stable materials is a long-standing challenge for the material design community. This task is difficult because stable materials only exist in a low-dimensional subspace of all possible periodic arrangements of atoms: 1) the coordinates must lie in the local energy minimum defined by quantum mechanics, and 2) global stability also requires the structure to follow the complex, yet specific bonding preferences between different atom types. Existing methods fail to incorporate these factors and often lack proper invariances. We propose a Crystal Diffusion Variational Autoencoder (CDVAE) that captures the physical inductive bias of material stability. By learning from the data distribution of stable materials, the decoder generates materials in a diffusion process that moves atomic coordinates towards a lower energy state and updates atom types to satisfy bonding preferences between neighbors. Our model also explicitly encodes interactions across periodic boundaries and respects permutation, translation, rotation, and periodic invariances. We significantly outperform past methods in three tasks: 1) reconstructing the input structure, 2) generating valid, diverse, and realistic materials, and 3) generating materials that optimize a specific property. We also provide several standard datasets and evaluation metrics for the broader machine learning community.

Task Affinity with Maximum Bipartite Matching in Few-Shot Learning

- Cat Phuoc Le, Juncheng Dong, Mohammadreza Soltani, Vahid Tarokh
- abstract@open-review(Poster): We propose an asymmetric affinity score for representing the complexity of utilizing the knowledge of one task for learning another one. Our method is based on the maximum bipartite matching algorithm and utilizes the Fisher Information matrix. We provide theoretical analyses demonstrating that the proposed score is mathematically well-defined, and subsequently use the affinity score to propose a novel algorithm for the few-shot learning problem. In particular, using this score, we find relevant training data labels to the test data and leverage the discovered relevant data for episodically fine-tuning a few-shot model. Results on various few-shot benchmark datasets demonstrate the efficacy of the proposed approach by improving the classification accuracy over the state-of-the-art methods even when using smaller models.

Latent Image Animator: Learning to Animate Images via Latent Space Navigation

- Yaohui Wang, Di Yang, Francois Bremond, Antitza Dantcheva
- abstract@open-review(Poster): Due to the remarkable progress of deep generative models, animating images has become increasingly efficient, whereas associated results have become increasingly realistic. Current animation-approaches commonly exploit structure representation extracted from driving videos. Such structure representation is instrumental in transferring motion from driving videos to still images. However, such approaches fail in case the source image and driving video encompass large appearance variation. Moreover, the extraction of structure information requires additional modules that endow the animation-model with increased complexity. Deviating from such models, we here introduce the Latent Image Animator (LIA), a self-supervised autoencoder that evades need for structure representation. LIA is streamlined to animate images by linear navigation in the latent space. Specifically, motion in generated video is constructed by linear displacement of codes in the latent space. Towards this, we learn a set of orthogonal motion directions simultaneously, and use their linear combination, in order to represent any displacement in the latent space. Extensive quantitative and qualitative analysis suggests that our model systematically and significantly outperforms state-of-art methods on VoxCeleb, Taichi and TED-talk datasets w.r.t. generated quality.

Know Thyself: Transferable Visual Control Policies Through Robot-Awareness

- Edward S. Hu, Kun Huang, Oleh Rybkin, Dinesh Jayaraman
- abstract@open-review(Poster): Training visual control policies from scratch on a new robot typically requires generating large amounts of robot-specific data. How might we leverage data previously collected on another robot to reduce or even completely remove this need for robot-specific data? We propose a "robot-aware control" paradigm that achieves this by exploiting readily available knowledge about the robot. We then instantiate this in a robot-aware model-based RL policy by training modular dynamics models that couple a transferable, robot-aware world dynamics module with a robot-specific, potentially analytical, robot dynamics module. This also enables us to set up visual planning costs that separately consider the robot agent and the world. Our experiments on tabletop manipulation tasks with simulated and real robots demonstrate that these plug-in improvements dramatically boost the transferability of visual model-based RL policies, even permitting zero-shot transfer of visual manipulation skills onto new robots. Project website: https://www.seas.upenn.edu/~hued/rac

Node Feature Extraction by Self-Supervised Multi-scale Neighborhood Prediction

- Eli Chien, Wei-Cheng Chang, Cho-Jui Hsieh, Hsiang-Fu Yu, Jiong Zhang, Olgica Milenkovic, Inderjit S Dhillon
- abstract@open-review(Poster): Learning on graphs has attracted significant attention in the learning community due to numerous real-world applications. In particular, graph neural networks (GNNs), which take \emph{numerical} node features and graph structure as inputs, have been shown to achieve state-of-the-art performance on various graph-related learning tasks. Recent works exploring the correlation between numerical node features and graph structure via self-supervised learning have paved the way for further performance improvements of GNNs. However, methods used for extracting numerical node features from \emph{raw data} are still \emph{graph-agnostic} within standard GNN pipelines. This practice is sub-optimal as it prevents one from fully utilizing potential correlations between graph topology and node attributes. To mitigate this issue, we propose a new self-supervised learning framework, Graph Information Aided Node feature exTraction (GIANT). GIANT makes use of the eXtreme Multi-label Classification (XMC) formalism, which is crucial for fine-tuning the language model based on graph information, and scales to large datasets. We also provide a theoretical analysis that justifies the use of XMC over link prediction and motivates integrating XR-Transformers, a powerful method for solving XMC problems, into the GIANT framework. We demonstrate the superior performance of GIANT over the standard GNN pipeline on Open Graph Benchmark datasets: For example, we improve the accuracy of the top-ranked method GAMLP from 68.25% to 69.67%, SGC from 63.29% to 66.10% and MLP from 47.24% to 61.10% on the ogbn-papers100M dataset by leveraging GIANT.

Spherical Message Passing for 3D Molecular Graphs

- Yi Liu, Limei Wang, Meng Liu, Yuchao Lin, Xuan Zhang, Bora Oztekin, Shuiwang Ji
- abstract@open-review(Poster): We consider representation learning of 3D molecular graphs in which each atom is associated with a spatial position in 3D. This is an under-explored area of research, and a principled message passing framework is currently lacking. In this work, we conduct analyses in the spherical coordinate system (SCS) for the complete identification of 3D graph structures. Based on such observations, we propose the spherical message passing (SMP) as a novel and powerful scheme for 3D molecular learning. SMP dramatically reduces training complexity, enabling it to perform efficiently on large-scale molecules. In addition, SMP is capable of distinguishing almost all molecular structures, and the uncovered cases may not exist in practice. Based on meaningful physically-based representations of 3D information, we further propose the SphereNet for 3D molecular learning. Experimental results demonstrate that the use of meaningful 3D information in SphereNet leads to significant performance improvements in prediction tasks. Our results also demonstrate the advantages of SphereNet in terms of capability, efficiency, and scalability.

Fairness Guarantees under Demographic Shift

- Stephen Giguere, Blossom Metevier, Bruno Castro da Silva, Yuriy Brun, Philip S. Thomas, Scott Niekum
- abstract@open-review(Poster): Recent studies have demonstrated that using machine learning for social applications can lead to injustice in the form of racist, sexist, and otherwise unfair and discriminatory outcomes. To address this challenge, recent machine learning algorithms have been designed to limit the likelihood such unfair behaviors will occur. However, these approaches typically assume the data used for training is representative of what will be encountered once the model is deployed, thus limiting their usefulness. In particular, if certain subgroups of the population become more or less probable after the model is deployed (a phenomenon we call demographic shift), the fair-ness assurances provided by prior algorithms are often invalid. We consider the impact of demographic shift and present a class of algorithms, called Shifty algorithms, that provide high-confidence behavioral guarantees that hold under demographic shift. Shifty is the first technique of its kind and demonstrates an effective strategy for designing algorithms to overcome the challenges demographic shift poses. We evaluate Shifty-ttest, an implementation of Shifty based on Studentâ€™s ð ‘¡-test, and, using a real-world data set of university entrance exams and subsequent student success, show that the models output by our algorithm avoid unfair bias under demo-graphic shift, unlike existing methods. Our experiments demonstrate that our algorithmâ€™s high-confidence fairness guarantees are valid in practice and that our algorithm is an effective tool for training models that are fair when demographic shift occurs.

Fooling Explanations in Text Classifiers

- Adam Ivankay, Ivan Girardi, Chiara Marchiori, Pascal Frossard
- abstract@open-review(Poster): State-of-the-art text classification models are becoming increasingly reliant on deep neural networks (DNNs). Due to their black-box nature, faithful and robust explanation methods need to accompany classifiers for deployment in real-life scenarios. However, it has been shown that explanation methods in vision applications are susceptible to local, imperceptible perturbations that can significantly alter the explanations without changing the predicted classes. We show here that the existence of such perturbations extends to text classifiers as well. Specifically, we introduce TextExplanationFooler (TEF), a novel explanation attack algorithm that alters text input samples imperceptibly so that the outcome of widely-used explanation methods changes considerably while leaving classifier predictions unchanged. We evaluate the attribution robustness estimation performance of TEF on five text classification datasets, utilizing three DNN architectures and a transformer architecture for each dataset. By significantly decreasing the correlation between unchanged and perturbed input attributions, we show that all models and explanation methods are susceptible to TEF perturbations. Moreover, we evaluate how the perturbations transfer to other model architectures and attribution methods, finding better than random performance in scenarios where the exact attacked model and explanation method are unknown. Finally, we introduce a semi-universal attack that is able to compute fast, computationally light perturbations with no knowledge of the attacked classifier nor explanation method. Overall, our work shows that explanations in text classifiers are fragile and users need to carefully address their robustness before relying on them in critical applications.

On the Learning and Learnability of Quasimetrics

- Tongzhou Wang, Phillip Isola
- abstract@open-review(Poster): Our world is full of asymmetries. Gravity and wind can make reaching a place easier than coming back. Social artifacts such as genealogy charts and citation graphs are inherently directed. In reinforcement learning and control, optimal goal-reaching strategies are rarely reversible (symmetrical). Distance functions supported on these asymmetrical structures are called quasimetrics. Despite their common appearance, little research has been done on the learning of quasimetrics. Our theoretical analysis reveals that a common class of learning algorithms, including unconstrained multilayer perceptrons (MLPs), provably fails to learn a quasimetric consistent with training data. In contrast, our proposed Poisson Quasimetric Embedding (PQE) is the first quasimetric learning formulation that both is learnable with gradient-based optimization and enjoys strong performance guarantees. Experiments on random graphs, social graphs, and offline Q-learning demonstrate its effectiveness over many common baselines.

Learning Prototype-oriented Set Representations for Meta-Learning

- Dan dan Guo, Long Tian, Minghe Zhang, Mingyuan Zhou, Hongyuan Zha
- abstract@open-review(Poster): Learning from set-structured data is a fundamental problem that has recently attracted increasing attention, where a series of summary networks are introduced to deal with the set input. In fact, many meta-learning problems can be treated as set-input tasks. Most existing summary networks aim to design different architectures for the input set in order to enforce permutation invariance. However, scant attention has been paid to the common cases where different sets in a meta distribution are closely related and share certain statistical properties. Viewing each set as a distribution over a set of global prototypes, this paper provides a novel prototype-oriented optimal transport (POT) framework to improve existing summary networks. To learn the distribution over the global prototypes, we minimize its regularized optimal transport distance to the set empirical distribution over data points, providing a natural unsupervised way to improve the summary network. Since our plug-and-play framework can be applied to many meta learning problems, we further instantiate it to the cases of few-shot classification and implicit meta generative modeling. Extensive experiments demonstrate that our framework significantly improves the existing summary networks on learning more powerful summary statistics from sets and can be successfully integrated into metric-based few-shot classification and generative modeling applications, providing a promising tool for addressing set-input and meta-learning problems.

Embedded-model flows: Combining the inductive biases of model-free deep learning and explicit probabilistic modeling

- Gianluigi Silvestri, Emily Fertig, Dave Moore, Luca Ambrogioni
- abstract@open-review(Poster): Normalizing flows have shown great success as general-purpose density estimators. However, many real world applications require the use of domain-specific knowledge, which normalizing flows cannot readily incorporate. We propose embedded-model flows (EMF), which alternate general-purpose transformations with structured layers that embed domain-specific inductive biases. These layers are automatically constructed by converting user-specified differentiable probabilistic models into equivalent bijective transformations. We also introduce gated structured layers, which allow bypassing the parts of the models that fail to capture the statistics of the data. We demonstrate that EMFs can be used to induce desirable properties such as multimodality and continuity. Furthermore, we show that EMFs enable a high performance form of variational inference where the structure of the prior model is embedded in the variational architecture. In our experiments, we show that this approach outperforms a large number of alternative methods in common structured inference problems.

A Relational Intervention Approach for Unsupervised Dynamics Generalization in Model-Based Reinforcement Learning

- Jiaxian Guo, Mingming Gong, Dacheng Tao
- abstract@open-review(Poster): The generalization of model-based reinforcement learning (MBRL) methods to environments with unseen transition dynamics is an important yet challenging problem. Existing methods try to extract environment-specified information Z from past transition segments to make the dynamics prediction model generalizable to different dynamics. However, because environments are not labelled, the extracted information inevitably contains redundant information unrelated to the dynamics in transition segments and thus fails to maintain a crucial property of Z: Z should be similar in the same environment and dissimilar in different ones. As a result, the learned dynamics prediction function will deviate from the true one, which undermines the generalization ability. To tackle this problem, we introduce an interventional prediction module to estimate the probability of two estimated $\hat{z}_i, \hat{z}_j$ belonging to the same environment. Furthermore, by utilizing the Z's invariance within a single environment, a relational head is proposed to enforce the similarity between $\hat{{Z}}$ from the same environment. As a result, the redundant information will be reduced in $\hat{Z}$. We empirically show that $\hat{{Z}}$ estimated by our method enjoy less redundant information than previous methods, and such

$\hat{{Z}}$ can significantly reduce dynamics prediction errors and improve the performance of model-based RL methods on zero-shot new environments with unseen dynamics. The codes of this method are available at \url{https://github.com/CR-Gjx/RIA}.

Critical Points in Quantum Generative Models

- Eric Ricardo Anschuetz
- abstract@open-review(Poster): One of the most important properties of neural networks is the clustering of local minima of the loss function near the global minimum, enabling efficient training. Though generative models implemented on quantum computers are known to be more expressive than their traditional counterparts, it has empirically been observed that these models experience a transition in the quality of their local minima. Namely, below some critical number of parameters, all local minima are far from the global minimum in function value; above this critical parameter count, all local minima are good approximators of the global minimum. Furthermore, for a certain class of quantum generative models, this transition has empirically been observed to occur at parameter counts exponentially large in the problem size, meaning practical training of these models is out of reach. Here, we give the first proof of this transition in trainability, specializing to this latter class of quantum generative model. We use techniques inspired by those used to study the loss landscapes of classical neural networks. We also verify that our analytic results hold experimentally even at modest model sizes.

VOS: Learning What You Don't Know by Virtual Outlier Synthesis

- Xuefeng Du, Zhaoning Wang, Mu Cai, Yixuan Li
- abstract@open-review(Poster): Out-of-distribution (OOD) detection has received much attention lately due to its importance in the safe deployment of neural networks. 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. Previous approaches rely on real outlier datasets for model regularization, which can be costly and sometimes infeasible to obtain in practice. In this paper, we present VOS, a novel framework for OOD detection by adaptively synthesizing virtual outliers that can meaningfully regularize the model's decision boundary during training. Specifically, VOS samples virtual outliers from the low-likelihood region of the class-conditional distribution estimated in the feature space. Alongside, we introduce a novel unknown-aware training objective, which contrastively shapes the uncertainty space between the ID data and synthesized outlier data. VOS achieves competitive performance on both object detection and image classification models, reducing the FPR95 by up to 9.36% compared to the previous best method on object detectors. Code is available at https://github.com/deeplearning-wisc/vos.

Trust Region Policy Optimisation in Multi-Agent Reinforcement Learning

- Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, Yaodong Yang
- abstract@open-review(Poster): Trust region methods rigorously enabled reinforcement learning (RL) agents to learn monotonically improving policies, leading to superior performance on a variety of tasks. Unfortunately, when it comes to multi-agent reinforcement learning (MARL), the property of monotonic improvement may not simply apply; this is because agents, even in cooperative games, could have conflicting directions of policy updates. As a result, achieving a guaranteed improvement on the joint policy where each agent acts individually remains an open challenge. In this paper, we extend the theory of trust region learning to MARL. Central to our findings are the multi-agent advantage decomposition lemma and the sequential policy update scheme. Based on these, we develop Heterogeneous-Agent Trust Region Policy Optimisation (HATPRO) and Heterogeneous-Agent Proximal Policy Optimisation (HAPPO) algorithms. Unlike many existing MARL algorithms, HATRPO/HAPPO do not need agents to share parameters, nor do they need any restrictive assumptions on decomposibility of the joint value function. Most importantly, we justify in theory the monotonic improvement property of HATRPO/HAPPO. We evaluate the proposed methods on a series of Multi-Agent MuJoCo and StarCraftII tasks. Results show that HATRPO and HAPPO significantly outperform strong baselines such as IPPO, MAPPO and MADDPG on all tested tasks, thereby establishing a new state of the art.

Unsupervised Disentanglement with Tensor Product Representations on the Torus

- Michael Rotman, Amit Dekel, Shir Gur, Yaron Oz, Lior Wolf
- abstract@open-review(Poster): The current methods for learning representations with auto-encoders almost exclusively employ vectors as the latent representations. In this work, we propose to employ a tensor product structure for this purpose. This way, the obtained representations are naturally disentangled. In contrast to the conventional variations methods, which are targeted toward normally distributed features, the latent space in our representation is distributed uniformly over a set of unit circles. We argue that the torus structure of the latent space captures the generative factors effectively. We employ recent tools for measuring unsupervised disentanglement, and in an extensive set of experiments demonstrate the advantage of our method in terms of disentanglement, completeness, and informativeness. The code for our proposed method is available at https://github.com/rotmanmi/Unsupervised-Disentanglement-Torus.

Anomaly Detection for Tabular Data with Internal Contrastive Learning

- Tom Shenkar, Lior Wolf
- abstract@open-review(Poster): We consider the task of finding out-of-class samples in tabular data, where little can be assumed on the structure of the data. In order to capture the structure of the samples of the single training class, we learn mappings that maximize the mutual information between each sample and the part that is masked out. The mappings are learned by employing a contrastive loss, which considers only one sample at a time. Once learned, we can score a test sample by measuring whether the learned mappings lead to a small contrastive loss using the masked parts of this sample. Our experiments show that our method leads by a sizable accuracy gap in comparison to the literature and that the same default set of hyperparameters provides state-of-the-art results across benchmarks.

LIGS: Learnable Intrinsic-Reward Generation Selection for Multi-Agent Learning

- David Henry Mguni, Taher Jafferjee, Jianhong Wang, Nicolas Perez-Nieves, Oliver Slumbers, Feifei Tong, Yang Li, Jiangcheng Zhu, Yaodong Yang, Jun Wang
- abstract@open-review(Poster): Efficient exploration is important for reinforcement learners (RL) to achieve high rewards. In multi-agent systems, coordinated exploration and behaviour is critical for agents to jointly achieve optimal outcomes. In this paper, we introduce a new general framework for improving coordination and performance of multi-agent reinforcement learners (MARL). Our framework, named Learnable Intrinsic-Reward Generation Selection algorithm (LIGS) introduces an adaptive learner, Generator that observes the agents and learns to construct intrinsic rewards online that coordinate the agentsâ€™ joint exploration and joint behaviour. Using a novel combination of reinforcement learning (RL) and switching controls, LIGS determines the best states to learn to add intrinsic rewards which leads to a highly efficient learning process. LIGS can subdivide complex tasks making them easier to solve and enables systems of RL agents to quickly solve environments with sparse rewards. LIGS can seamlessly adopt existing multi-agent RL algorithms and our theory shows that it ensures convergence to joint policies that deliver higher system performance. We demonstrate the superior performance of the LIGS framework in challenging tasks in Foraging and StarCraft II and show LIGS is capable of tackling tasks previously unsolvable by MARL methods.

Bayesian Modeling and Uncertainty Quantification for Learning to Optimize: What, Why, and How

- Yuning You, Yue Cao, Tianlong Chen, Zhangyang Wang, Yang Shen

- abstract@open-review(Poster): Optimizing an objective function with uncertainty awareness is well-known to improve the accuracy and confidence of optimization solutions. Meanwhile, another relevant but very different question remains yet open: how to model and quantify the uncertainty of an optimization algorithm (a.k.a., optimizer) itself? To close such a gap, the prerequisite is to consider the optimizers as sampled from a distribution, rather than a few prefabricated and fixed update rules. We first take the novel angle to consider the algorithmic space of optimizers, and provide definitions for the optimizer prior and likelihood, that intrinsically determine the posterior and therefore uncertainty. We then leverage the recent advance of learning to optimize (L2O) for the space parameterization, with the end-to-end training pipeline built via variational inference, referred to as uncertainty-aware L2O (UA-L2O). Our study represents the first effort to recognize and quantify the uncertainty of the optimization algorithm. The extensive numerical results show that, UA-L2O achieves superior uncertainty calibration with accurate confidence estimation and tight confidence intervals, suggesting the improved posterior estimation thanks to considering optimizer uncertainty. Intriguingly, UA-L2O even improves optimization performances for two out of three test functions, the loss function in data privacy attack, and four of five cases of the energy function in protein docking. Our codes are released at https://github.com/Shen-Lab/Bayesian-L2O.

Deep Ensembling with No Overhead for either Training or Testing: The All-Round Blessings of Dynamic Sparsity

- Shiwei Liu, Tianlong Chen, Zahra Atashgahi, Xiaohan Chen, Ghada Sokar, Elena Mocanu, Mykola Pechenizkiy, Zhangyang Wang, Decebal Constantin Mocanu
- abstract@open-review(Poster): The success of deep ensembles on improving predictive performance, uncertainty estimation, and out-of-distribution robustness has been extensively studied in the machine learning literature. Albeit the promising results, naively training multiple deep neural networks and combining their predictions at inference leads to prohibitive computational costs and memory requirements. Recently proposed efficient ensemble approaches reach the performance of the traditional deep ensembles with significantly lower costs. However, the training resources required by these approaches are still at least the same as training a single dense model. In this work, we draw a unique connection between sparse neural network training and deep ensembles, yielding a novel efficient ensemble learning framework called $FreeTickets$. Instead of training multiple dense networks and averaging them, we directly train sparse subnetworks from scratch and extract diverse yet accurate subnetworks during this efficient, sparse-to-sparse training. Our framework, $FreeTickets$, is defined as the ensemble of these relatively cheap sparse subnetworks. Despite being an ensemble method, $FreeTickets$ has even fewer parameters and training FLOPs than a single dense model. This seemingly counter-intuitive outcome is due to the ultra training/inference efficiency of dynamic sparse training. $FreeTickets$ surpasses the dense baseline in all the following criteria: prediction accuracy, uncertainty estimation, out-of-distribution (OoD) robustness, as well as efficiency for both training and inference. Impressively, $FreeTickets$ outperforms the naive deep ensemble with ResNet50 on ImageNet using around only $1/5$ of the training FLOPs required by the latter. We have released our source code at https://github.com/VITA-Group/FreeTickets.

HyperDQN: A Randomized Exploration Method for Deep Reinforcement Learning

- Ziniu Li, Yingru Li, Yushun Zhang, Tong Zhang, Zhi-Quan Luo
- abstract@open-review(Poster): Randomized least-square value iteration (RLSVI) is a provably efficient exploration method. However, it is limited to the case where (1) a good feature is known in advance and (2) this feature is fixed during the training. If otherwise, RLSVI suffers an unbearable computational burden to obtain the posterior samples. In this work, we present a practical algorithm named HyperDQN to address the above issues under deep RL. In addition to a non-linear neural network (i.e., base model) that predicts Q-values, our method employs a probabilistic hypermodel (i.e., meta model), which outputs the parameter of the base model. When both models are jointly optimized under a specifically designed objective, three purposes can be achieved. First, the hypermodel can generate approximate posterior samples regarding the parameter of the Q-value function. As a result, diverse Q-value functions are sampled to select exploratory action sequences. This retains the punchline of RLSVI for efficient exploration. Second, a good feature is learned to approximate Q-value functions. This addresses limitation (1). Third, the posterior samples of the Q-value function can be obtained in a more efficient way than the existing methods, and the changing feature does not affect the efficiency. This deals with limitation (2). On the Atari suite, HyperDQN with 20M frames outperforms DQN with 200M frames in terms of the maximum human-normalized score. For SuperMarioBros, HyperDQN outperforms several exploration bonus and randomized exploration methods on 5 out of 9 games.

Unraveling Model-Agnostic Meta-Learning via The Adaptation Learning Rate

- Yingtian Zou, Fusheng Liu, Qianxiao Li
- abstract@open-review(Poster): Model-Agnostic Meta-Learning (MAML) aims to find initial weights that allow fast adaptation to new tasks. The adaptation (inner loop) learning rate in MAML plays a central role in enabling such fast adaptation. However, how to choose this value in practice and how this choice affects the adaptation error remains less explored. In this paper, we study the effect of the adaptation learning rate in meta-learning with mixed linear regression. First, we present a principled way to estimate optimal adaptation learning rates that minimize the population risk of MAML. Second, we interpret the underlying dependence between the optimal adaptation learning rate and the input data. Finally, we prove that compared with empirical risk minimization (ERM), MAML produces an initialization with a smaller average distance to the task optima, consistent with previous practical findings. These results are corroborated with numerical experiments.

iFlood: A Stable and Effective Regularizer

- Yuexiang Xie, Zhen WANG, Yaliang Li, Ce Zhang, Jingren Zhou, Bolin Ding
- abstract@open-review(Poster): Various regularization methods have been designed to prevent overfitting of machine learning models. Among them, a surprisingly simple yet effective one, called Flooding, is proposed recently, which directly constrains the training loss on average to stay at a given level. However, our further studies uncover that the design of the loss function of Flooding can lead to a discrepancy between its objective and implementation, and cause the instability issue. To resolve these issues, in this paper, we propose a new regularizer, called individual Flood (denoted as iFlood). With instance-level constraints on training loss, iFlood encourages the trained models to better fit the under-fitted instances while suppressing the confidence on over-fitted ones. We theoretically show that the design of iFlood can be intrinsically connected with removing the noise or bias in training data, which makes it suitable for a variety of applications to improve the generalization performances of learned models. We also theoretically link iFlood to some other regularizers by comparing the inductive biases they introduce. Our experimental results on both image classification and language understanding tasks confirm that models learned with iFlood can stably converge to solutions with better generalization ability, and behave consistently at instance-level.

FlexConv: Continuous Kernel Convolutions With Differentiable Kernel Sizes

- David W. Romero, Robert-Jan Bruintjes, Jakub Mikolaj Tomczak, Erik J Bekkers, Mark Hoogendoorn, Jan van Gemert
- abstract@open-review(Poster): When designing Convolutional Neural Networks (CNNs), one must select the size of the convolutional kernels before training. Recent works show CNNs benefit from different kernel sizes at different layers, but exploring all possible combinations is unfeasible in practice. A more efficient approach is to learn the kernel size during training. However, existing works that learn the kernel size have a limited bandwidth. These approaches scale kernels by dilation, and thus the detail they can describe is limited. In this work, we propose FlexConv, a novel convolutional operation with which high bandwidth convolutional kernels of learnable kernel size can be learned at a fixed parameter cost. FlexNets model long-term dependencies without the use of pooling, achieve state-of-the-art performance on several sequential datasets, outperform recent works with learned kernel sizes, and are competitive with much deeper ResNets on image benchmark datasets. Additionally, FlexNets can be deployed at higher resolutions than those seen during training. To avoid aliasing, we propose a novel kernel parameterization with which the frequency of the kernels can be analytically controlled. Our novel kernel parameterization shows higher descriptive power and faster convergence speed than existing parameterizations. This leads to important improvements in classification accuracy.

Zero Pixel Directional Boundary by Vector Transform

- Edoardo Mello Rella, Ajad Chhatkuli, Yun Liu, Ender Konukoglu, Luc Van Gool
- abstract@open-review(Poster): Boundaries or contours are among the primary visual cues used by human and computer vision systems. One of the key problems in boundary detection is the loss formulation, which typically leads to class imbalance and, as a consequence, to thick boundaries which require non-differential post-processing steps to be thinned. In this paper, we re-interpret boundaries as 1-D surfaces and formulate a one-to-one vector transform function that allows for training of boundary prediction completely avoiding the class imbalance issue. Specifically, we define the boundary representation at any point as the unit vector pointing to the closest boundary surface. Our problem formulation leads to the estimation of direction as well as richer contextual information of the boundary, and, if desired, the availability of zero-pixel thin boundaries also at training time. Our method uses no hyper-parameter in the training loss and a fixed stable hyper-parameter at inference. We provide theoretical justification/discussions of the vector transform representation. We evaluate the proposed loss method using a standard architecture and show the excellent performance over other losses and representations on several datasets.

A Conditional Point Diffusion-Refinement Paradigm for 3D Point Cloud Completion

- Zhaoyang Lyu, Zhifeng Kong, Xudong XU, Liang Pan, Dahua Lin
- abstract@open-review(Poster): 3D point clouds are an important data format that captures 3D information for real world objects. Since 3D point clouds scanned in the real world are often incomplete, it is important to recover the complete point cloud for many downstreaming applications. Most existing point cloud completion methods use the Chamfer Distance (CD) loss for training. The CD loss estimates correspondences between two point clouds by searching nearest neighbors, which does not capture the overall point distribution on the generated shape, and therefore likely leads to non-uniform point cloud generation. To tackle this problem, we propose a novel Point Diffusion-Refinement (PDR) paradigm for point cloud completion. PDR consists of a Conditional Generation Network (CGNet) and a ReFinement Network (RFNet). The CGNet uses a conditional generative model called the denoising diffusion probabilistic model (DDPM) to generate a coarse completion conditioned on the partial observation. DDPM establishes a one-to-one pointwise mapping between the generated point cloud and the uniform ground truth, and then optimizes the mean squared error loss to realize uniform generation. The RFNet refines the coarse output of the CGNet and further improves quality of the completed point cloud. In terms of the architecture, we develop a novel dual-path architecture for both networks. The architecture can (1) effectively and efficiently extract multi-level features from partially observed point clouds to guide completion, and (2) accurately manipulate spatial locations of 3D points to obtain smooth surfaces and sharp details. Extensive experimental results on various benchmark datasets show that our PDR paradigm outperforms previous state-of-the-art methods for point cloud completion. In addition, with the help of the RFNet, we can accelerate the iterative generation process of the DDPM by up to 50 times without much performance drop.

Auto-Transfer: Learning to Route Transferable Representations

- Keerthiram Murugesan, Vijay Sadashivaiah, Ronny Luss, Karthikeyan Shanmugam, Pin-Yu Chen, Amit Dhurandhar
- abstract@open-review(Poster): Knowledge transfer between heterogeneous source and target networks and tasks has received a lot of attention in recent times as large amounts of quality labeled data can be difficult to obtain in many applications. Existing approaches typically constrain the target deep neural network (DNN) feature representations to be close to the source DNNs feature representations, which can be limiting. We, in this paper, propose a novel adversarial multi-armed bandit approach that automatically learns to route source representations to appropriate target representations following which they are combined in meaningful ways to produce accurate target models. We see upwards of 5\% accuracy improvements compared with the state-of-the-art knowledge transfer methods on four benchmark (target) image datasets CUB200, Stanford Dogs, MIT67, and Stanford40 where the source dataset is ImageNet. We qualitatively analyze the goodness of our transfer scheme by showing individual examples of the important features focused on by our target network at different layers compared with the (closest) competitors. We also observe that our improvement over other methods is higher for smaller target datasets making it an effective tool for small data applications that may benefit from transfer learning.

PoNet: Pooling Network for Efficient Token Mixing in Long Sequences

- Chao-Hong Tan, Qian Chen, Wen Wang, Qinglin Zhang, Siqi Zheng, Zhen-Hua Ling
- abstract@open-review(Poster): Transformer-based models have achieved great success in various NLP, vision, and speech tasks. However, the core of Transformer, the self-attention mechanism, has a quadratic time and memory complexity with respect to the sequence length, which hinders applications of Transformer-based models to long sequences. Many approaches have been proposed to mitigate this problem, such as sparse attention mechanisms, low-rank matrix approximations and scalable kernels, and token mixing alternatives to self-attention. We propose a novel Pooling Network (PoNet) for token mixing in long sequences with linear complexity. We design multi-granularity pooling and pooling fusion to capture different levels of contextual information and combine their interactions with tokens. On the Long Range Arena benchmark, PoNet significantly outperforms Transformer and achieves competitive accuracy, while being only slightly slower than the fastest model, FNet, across all sequence lengths measured on GPUs. We also conduct systematic studies on the transfer learning capability of PoNet and observe that PoNet achieves 95.7 percent of the accuracy of BERT on the GLUE benchmark, outperforming FNet by 4.5 percent relative. Comprehensive ablation analysis demonstrates effectiveness of the designed multi-granularity pooling and pooling fusion for token mixing in long sequences and efficacy of the designed pre-training tasks for PoNet to learn transferable contextualized language representations.

Huber Additive Models for Non-stationary Time Series Analysis

- Yingjie Wang, Xianrui Zhong, Fengxiang He, Hong Chen, Dacheng Tao
- abstract@open-review(Poster): Sparse additive models have shown promising ï¬‚exibility and interpretability in processing time series data. However, existing methods usually assume the time series data to be stationary and the innovation is sampled from a Gaussian distribution. Both assumptions are too stringent for heavy-tailed and non-stationary time series data that frequently arise in practice, such as ï¬ nance and medical ï¬ elds. To address these problems, we propose an adaptive sparse Huber additive model for robust forecasting in both non-Gaussian data and (non)stationary data. In theory, the generalization bounds of our estimator are established for both stationary and nonstationary time series data, which are independent of the widely used mixing conditions in learning theory of dependent observations. Moreover, the error bound for non-stationary time series contains a discrepancy measure for the shifts of the data distributions over time. Such a discrepancy measure can be estimated empirically and used as a penalty in our method. Experimental results on both synthetic and real-world benchmark datasets validate the effectiveness of the proposed method. The code is available at https://github.com/xianruizhong/SpHAM.

Model-augmented Prioritized Experience Replay

- Youngmin Oh, Jinwoo Shin, Eunho Yang, Sung Ju Hwang
- abstract@open-review(Poster): Experience replay is an essential component in off-policy model-free reinforcement learning (MfRL). Due to its effectiveness, various methods for calculating priority scores on experiences have been proposed for sampling. Since critic networks are crucial to policy learning, TD-error, directly correlated to Q-values, is one of the most frequently used features to compute the scores. However, critic networks often under- or overestimate Q-values, so it is often ineffective to learn to predict Q-values by sampled experiences based heavily on TD-error. Accordingly, it is valuable to find auxiliary features, which positively support TD-error in calculating the scores for efficient sampling. Motivated by this, we propose a novel experience replay method, which we call model-augmented prioritized experience replay (MaPER), that employs new learnable features driven from components in model-based RL (MbRL) to calculate the scores on experiences. The proposed MaPER brings the effect of curriculum

learning for predicting Q-values better by the critic network with negligible memory and computational overhead compared to the vanilla PER. Indeed, our experimental results on various tasks demonstrate that MaPER can significantly improve the performance of the state-of-the-art off-policy MfRL and MbRL which includes off-policy MfRL algorithms in its policy optimization procedure.

Post-Training Detection of Backdoor Attacks for Two-Class and Multi-Attack Scenarios

- Zhen Xiang, David Miller, George Kesidis
- abstract@open-review(Poster): Backdoor attacks (BAs) are an emerging threat to deep neural network classifiers. A victim classifier will predict to an attacker-desired target class whenever a test sample is embedded with the same backdoor pattern (BP) that was used to poison the classifier's training set. Detecting whether a classifier is backdoor attacked is not easy in practice, especially when the defender is, e.g., a downstream user without access to the classifier's training set. This challenge is addressed here by a reverse-engineering defense (RED), which has been shown to yield state-of-the-art performance in several domains. However, existing REDs are not applicable when there are only two classes or when multiple attacks are present. These scenarios are first studied in the current paper, under the practical constraints that the defender neither has access to the classifier's training set nor to supervision from clean reference classifiers trained for the same domain. We propose a detection framework based on BP reverse-engineering and a novel expected transferability (ET) statistic. We show that our ET statistic is effective using the same detection threshold, irrespective of the classification domain, the attack configuration, and the BP reverse-engineering algorithm that is used. The excellent performance of our method is demonstrated on six benchmark datasets. Notably, our detection framework is also applicable to multi-class scenarios with multiple attacks. Code is available at https://github.com/zhenxianglance/2ClassBADetection.

Multi-Task Processes

- Donggyun Kim, Seongwoong Cho, Wonkwang Lee, Seunghoon Hong
- abstract@open-review(Poster): Neural Processes (NPs) consider a task as a function realized from a stochastic process and flexibly adapt to unseen tasks through inference on functions. However, naive NPs can model data from only a single stochastic process and are designed to infer each task independently. Since many real-world data represent a set of correlated tasks from multiple sources (e.g., multiple attributes and multi-sensor data), it is beneficial to infer them jointly and exploit the underlying correlation to improve the predictive performance. To this end, we propose Multi-Task Neural Processes (MTNPs), an extension of NPs designed to jointly infer tasks realized from multiple stochastic processes. We build MTNPs in a hierarchical way such that inter-task correlation is considered by conditioning all per-task latent variables on a single global latent variable. In addition, we further design our MTNPs so that they can address multi-task settings with incomplete data (i.e., not all tasks share the same set of input points), which has high practical demands in various applications. Experiments demonstrate that MTNPs can successfully model multiple tasks jointly by discovering and exploiting their correlations in various real-world data such as time series of weather attributes and pixel-aligned visual modalities. We release our code at https://github.com/GitGyun/multi_task_neural_processes.

Dynamic Token Normalization improves Vision Transformers

- Wenqi Shao, Yixiao Ge, Zhaoyang Zhang, XUYUAN XU, Xiaogang Wang, Ying Shan, Ping Luo
- abstract@open-review(Poster): Vision Transformer (ViT) and its variants (e.g., Swin, PVT) have achieved great success in various computer vision tasks, owing to their capability to learn long-range contextual information. Layer Normalization (LN) is an essential ingredient in these models. However, we found that the ordinary LN makes tokens at different positions similar in magnitude because it normalizes embeddings within each token. It is difficult for Transformers to capture inductive bias such as the positional context in an image with LN. We tackle this problem by proposing a new normalizer, termed Dynamic Token Normalization (DTN), where normalization is performed both within each token (intra-token) and across different tokens (inter-token). DTN has several merits. Firstly, it is built on a unified formulation and thus can represent various existing normalization methods. Secondly, DTN learns to normalize tokens in both intra-token and inter-token manners, enabling Transformers to capture both the global contextual information and the local positional context. Thirdly, by simply replacing LN layers, DTN can be readily plugged into various vision transformers, such as ViT, Swin, and PVT. Extensive experiments show that the transformer equipped with DTN consistently outperforms baseline model with minimal extra parameters and computational overhead. For example, DTN outperforms LN on small ViT by 1.1% top-1 accuracy on ImageNet.

Symbolic Learning to Optimize: Towards Interpretability and Scalability

- Wenqing Zheng, Tianlong Chen, Ting-Kuei Hu, Zhangyang Wang
- abstract@open-review(Poster): Recent studies on Learning to Optimize (L2O) suggest a promising path to automating and accelerating the optimization procedure for complicated tasks. Existing L2O models parameterize optimization rules by neural networks, and learn those numerical rules via meta-training. However, they face two common pitfalls: (1) scalability: the numerical rules represented by neural networks create extra memory overhead for applying L2O models, and limits their applicability to optimizing larger tasks; (2) interpretability: it is unclear what each L2O model has learned in its black-box optimization rule, nor is it straightforward to compare different L2O models in an explainable way. To avoid both pitfalls, this paper proves the concept that we can "kill two birds by one stone", by introducing the powerful tool of symbolic regression to L2O. In this paper, we establish a holistic symbolic representation and analysis framework for L2O, which yields a series of insights for learnable optimizers. Leveraging our findings, we further propose a lightweight L2O model that can be meta-trained on large-scale problems and outperformed human-designed and tuned optimizers. Our work is set to supply a brand-new perspective to L2O research. Codes are available at: https://github.com/VITA-Group/Symbolic-Learning-To-Optimize.

Sequential Reptile: Inter-Task Gradient Alignment for Multilingual Learning

- Seanie Lee, Hae Beom Lee, Juho Lee, Sung Ju Hwang
- abstract@open-review(Poster): Multilingual models jointly pretrained on multiple languages have achieved remarkable performance on various multilingual downstream tasks. Moreover, models finetuned on a single monolingual downstream task have shown to generalize to unseen languages. In this paper, we first show that it is crucial for those tasks to align gradients between them in order to maximize knowledge transfer while minimizing negative transfer. Despite its importance, the existing methods for gradient alignment either have a completely different purpose, ignore inter-task alignment, or aim to solve continual learning problems in rather inefficient ways. As a result of the misaligned gradients between tasks, the model suffers from severe negative transfer in the form of catastrophic forgetting of the knowledge acquired from the pretraining. To overcome the limitations, we propose a simple yet effective method that can efficiently align gradients between tasks. Specifically, we perform each inner-optimization by sequentially sampling batches from all the tasks, followed by a Reptile outer update. Thanks to the gradients aligned between tasks by our method, the model becomes less vulnerable to negative transfer and catastrophic forgetting. We extensively validate our method on various multi-task learning and zero-shot cross-lingual transfer tasks, where our method largely outperforms all the relevant baselines we consider.

Pseudo Numerical Methods for Diffusion Models on Manifolds

- Luping Liu, Yi Ren, Zhijie Lin, Zhou Zhao
- abstract@open-review(Poster): Denoising Diffusion Probabilistic Models (DDPMs) can generate high-quality samples such as image and audio samples. However, DDPMs require hundreds to thousands of iterations to produce a sample. Several prior works have successfully accelerated DDPMs through adjusting the variance schedule (e.g., Improved Denoising Diffusion Probabilistic Models) or the denoising equation (e.g., Denoising Diffusion Implicit Models (DDIMs)). However, these acceleration methods cannot maintain the quality of samples and even introduce new noise at high speedup rate, which limit their practicability. To accelerate the inference process while keeping the sample quality, we provide a new perspective that DDPMs should be

treated as solving differential equations on manifolds. Under such a perspective, we propose pseudo numerical methods for diffusion models (PNDMs). Specifically, we figure out how to solve differential equations on manifolds and show that DDIMs are simple cases of pseudo numerical methods. We change several classical numerical methods to corresponding pseudo numerical methods and find that pseudo linear multi-step method is the best method in most situations. According to our experiments, by directly using pre-trained models on Cifar10, CelebA and LSUN, PNDMs can generate higher quality synthetic images with only 50 steps compared with 1000-step DDIMs (20x speedup), significantly outperform DDIMs with 250 steps (by around 0.4 in FID) and have good generalization on different variance schedules.

Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm

- Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, Junjie Yan
- abstract@open-review(Poster): Recently, large-scale Contrastive Language-Image Pre-training (CLIP) has attracted unprecedented attention for its impressive zero-shot recognition ability and excellent transferability to downstream tasks. However, CLIP is quite data-hungry and requires 400M image-text pairs for pre-training, thereby restricting its adoption. This work proposes a novel training paradigm, Data efficient CLIP (DeCLIP), to alleviate this limitation. We demonstrate that by carefully utilizing the widespread supervision among the image-text pairs, our De-CLIP can learn generic visual features more efficiently. Instead of using the single image-text contrastive supervision, we fully exploit data potential through the use of (1) self-supervision within each modality; (2) multi-view supervision across modalities; (3) nearest-neighbor supervision from other similar pairs. Benefiting from intrinsic supervision, our DeCLIP-ResNet50 can achieve 60.4% zero-shot top1 accuracy on ImageNet, which is 0.8% above the CLIP-ResNet50 while using 7.1Ã—fewer data. Our DeCLIP-ResNet50 outperforms its counterpart in 8 out of 11 visual datasets when transferred to downstream tasks. Moreover, Scaling up the model and computing also works well in our framework.

Environment Predictive Coding for Visual Navigation

- Santhosh Kumar Ramakrishnan, Tushar Nagarajan, Ziad Al-Halah, Kristen Grauman
- abstract@open-review(Poster): We introduce environment predictive coding, a self-supervised approach to learn environment-level representations for embodied agents. In contrast to prior work on self-supervised learning for individual images, we aim to encode a 3D environment using a series of images observed by an agent moving in it. We learn these representations via a masked-zone prediction task, which segments an agentâ€™s trajectory into zones and then predicts features of randomly masked zones, conditioned on the agentâ€™s camera poses. This explicit spatial conditioning encourages learning representations that capture the geometric and semantic regularities of 3D environments. We learn such representations on a collection of video walkthroughs and demonstrate successful transfer to multiple downstream navigation tasks. Our experiments on the real-world scanned 3D environments of Gibson and Matterport3D show that our method obtains 2 - 6Ã— higher sample-efï¬ ciency and up to 57% higher performance over standard image-representation learning.

Topological Experience Replay

- Zhang-Wei Hong, Tao Chen, Yen-Chen Lin, Joni Pajarinen, Pulkit Agrawal
- abstract@open-review(Poster): State-of-the-art deep Q-learning methods update Q-values using state transition tuples sampled from the experience replay buffer. This strategy often randomly samples or prioritizes data sampling based on measures such as the temporal difference (TD) error. Such sampling strategies can be inefficient at learning Q-function since a state's correct Q-value preconditions on the accurate successor states' Q-value. Disregarding such a successor's value dependency leads to useless updates and even learning wrong values. To expedite Q-learning, we maintain states' dependency by organizing the agent's experience into a graph. Each edge in the graph represents a transition between two connected states. We perform value backups via a breadth-first search that expands vertices in the graph starting from the set of terminal states successively moving backward. We empirically show that our method is substantially more data-efficient than several baselines on a diverse range of goal-reaching tasks. Notably, the proposed method also outperforms baselines that consume more batches of training experience.

Sparsity Winning Twice: Better Robust Generalization from More Efficient Training

- Tianlong Chen, Zhenyu Zhang, pengjun wang, Santosh Balachandra, Haoyu Ma, Zehao Wang, Zhangyang Wang
- abstract@open-review(Poster): Recent studies demonstrate the deep networks, even robustified by the state-of-the-art adversarial training (AT), still suffer from large robust generalization gaps, in addition to the much more expensive training costs than standard training. In this paper, we investigate this intriguing problem from a new perspective, i.e., $\textit{injecting appropriate forms of sparsity}$ during adversarial training. We introduce two alternatives for sparse adversarial training: (i) $\textit{static sparsity}$, by leveraging recent results from the lottery ticket hypothesis to identify critical sparse subnetworks arising from the early training; (ii) $\textit{dynamic sparsity}$, by allowing the sparse subnetwork to adaptively adjust its connectivity pattern (while sticking to the same sparsity ratio) throughout training. We find both static and dynamic sparse methods to yield win-win: substantially shrinking the robust generalization gap and alleviating the robust overfitting, meanwhile significantly saving training and inference FLOPs. Extensive experiments validate our proposals with multiple network architectures on diverse datasets, including CIFAR-10/100 and Tiny-ImageNet. For example, our methods reduce robust generalization gap and overfitting by 34.44% and 4.02%, with comparable robust/standard accuracy boosts and 87.83%/87.82% training/inference FLOPs savings on CIFAR-100 with ResNet-18. Besides, our approaches can be organically combined with existing regularizers, establishing new state-of-the-art results in AT. All codes are included.

CrossMatch: Cross-Classifier Consistency Regularization for Open-Set Single Domain Generalization

- Ronghang Zhu, Sheng Li
- abstract@open-review(Poster): Single domain generalization (SDG) is a challenging scenario of domain generalization, where only one source domain is available to train the model. Typical SDG methods are based on the adversarial data augmentation strategy, which complements the diversity of source domain to learn a robust model. Existing SDG methods require the source and target domains to have the same label space. However, as target domains may contain novel categories unseen in source label space, this assumption is not practical in many real-world applications. In this paper, we propose a challenging and untouched problem: \textit{Open-Set Single Domain Generalization} (OS-SDG), where target domains include unseen categories out of source label space. The goal of OS-SDG is to learn a model, with only one source domain, to classify a target sample with correct class if it belongs to source label space, or assign it to unknown classes. We design a \textit{CrossMatch} approach to improve the performance of SDG methods on identifying unknown classes by leveraging a multi-binary classifier. CrossMatch generates auxiliary samples out of source label space by using an adversarial data augmentation strategy. We also adopt a consistency regularization on generated auxiliary samples between multi-binary classifiers and the model trained by SDG methods, to improve the modelâ€™s capability on unknown class identification. Experimental results on benchmark datasets prove the effectiveness of CrossMatch on enhancing the performance of SDG methods in the OS-SDG setting.

Robust Unlearnable Examples: Protecting Data Privacy Against Adversarial Learning

- Shaopeng Fu, Fengxiang He, Yang Liu, Li Shen, Dacheng Tao
- abstract@open-review(Poster): The tremendous amount of accessible data in cyberspace face the risk of being unauthorized used for training deep learning models. To address this concern, methods are proposed to make data unlearnable for deep learning models by adding a type of error-minimizing noise. However, such conferred unlearnability is found fragile to adversarial training. In this paper, we design new methods to generate robust unlearnable examples that are protected from adversarial training. We first find that the vanilla error-minimizing noise, which suppresses the informative knowledge of data via minimizing the corresponding training loss, could not effectively minimize the adversarial training loss. This explains the vulnerability of error-

minimizing noise in adversarial training. Based on the observation, robust error-minimizing noise is then introduced to reduce the adversarial training loss. Experiments show that the unlearnability brought by robust error-minimizing noise can effectively protect data from adversarial training in various scenarios. The code is available at \url{https://github.com/fshp971/robust-unlearnable-examples}.

A NON-PARAMETRIC REGRESSION VIEWPOINT : GENERALIZATION OF OVERPARAMETRIZED DEEP RELU NETWORK UNDER NOISY OBSERVATIONS

- Namjoon Suh, Hyunouk Ko, Xiaoming Huo
- abstract@open-review(Poster): We study the generalization properties of the overparameterized deep neural network (DNN) with Rectified Linear Unit (ReLU) activations. Under the non-parametric regression framework, it is assumed that the ground-truth function is from a reproducing kernel Hilbert space (RKHS) induced by a neural tangent kernel (NTK) of ReLU DNN, and a dataset is given with the noises. Without a delicate adoption of early stopping, we prove that the overparametrized DNN trained by vanilla gradient descent does not recover the ground-truth function. It turns out that the estimated DNN's L_{2} prediction error is bounded away from 0. As a complement of the above result, we show that the ℓ_{2}-regularized gradient descent enables the overparametrized DNN achieve the minimax optimal convergence rate of the L_{2} prediction error, without early stopping. Notably, the rate we obtained is faster than $\mathcal{O}(n^{-1/2})$ known in the literature.

Active Hierarchical Exploration with Stable Subgoal Representation Learning

- Siyuan Li, Jin Zhang, Jianhao Wang, Yang Yu, Chongjie Zhang
- abstract@open-review(Poster): Goal-conditioned hierarchical reinforcement learning (GCHRL) provides a promising approach to solving long-horizon tasks. Recently, its success has been extended to more general settings by concurrently learning hierarchical policies and subgoal representations. Although GCHRL possesses superior exploration ability by decomposing tasks via subgoals, existing GCHRL methods struggle in temporally extended tasks with sparse external rewards, since the high-level policy learning relies on external rewards. As the high-level policy selects subgoals in an online learned representation space, the dynamic change of the subgoal space severely hinders effective high-level exploration. In this paper, we propose a novel regularization that contributes to both stable and efficient subgoal representation learning. Building upon the stable representation, we design measures of novelty and potential for subgoals, and develop an active hierarchical exploration strategy that seeks out new promising subgoals and states without intrinsic rewards. Experimental results show that our approach significantly outperforms state-of-the-art baselines in continuous control tasks with sparse rewards.

Deep AutoAugment

- Yu Zheng, Zhi Zhang, Shen Yan, Mi Zhang
- abstract@open-review(Poster): While recent automated data augmentation methods lead to state-of-the-art results, their design spaces and the derived data augmentation strategies still incorporate strong human priors. In this work, instead of fixing a set of hand-picked default augmentations alongside the searched data augmentations, we propose a fully automated approach for data augmentation search named Deep AutoAugment (DeepAA). DeepAA progressively builds a multi-layer data augmentation pipeline from scratch by stacking augmentation layers one at a time until reaching convergence. For each augmentation layer, the policy is optimized to maximize the cosine similarity between the gradients of the original and augmented data along the direction with low variance. Our experiments show that even without default augmentations, we can learn an augmentation policy that achieves strong performance with that of previous works. Extensive ablation studies show that the regularized gradient matching is an effective search method for data augmentation policies. Our code is available at: https://github.com/MSU-MLSys-Lab/DeepAA .

Temporal Alignment Prediction for Supervised Representation Learning and Few-Shot Sequence Classification

- Bing Su, Ji-Rong Wen
- abstract@open-review(Poster): Explainable distances for sequence data depend on temporal alignment to tackle sequences with different lengths and local variances. Most sequence alignment methods infer the optimal alignment by solving an optimization problem under pre-defined feasible alignment constraints, which not only is time-consuming, but also makes end-to-end sequence learning intractable. In this paper, we propose a learnable sequence distance called Temporal Alignment Prediction (TAP). TAP employs a lightweight convolutional neural network to directly predict the optimal alignment between two sequences, so that only straightforward calculations are required and no optimization is involved in inference. TAP can be applied in different distance-based machine learning tasks. For supervised sequence representation learning, we show that TAP trained with various metric learning losses achieves completive performances with much faster inference speed. For few-shot action classification, we apply TAP as the distance measure in the metric learning-based episode-training paradigm. This simple strategy achieves comparable results with state-of-the-art few-shot action recognition methods.

Anti-Oversmoothing in Deep Vision Transformers via the Fourier Domain Analysis: From Theory to Practice

- Peihao Wang, Wenqing Zheng, Tianlong Chen, Zhangyang Wang
- abstract@open-review(Poster): Vision Transformer (ViT) has recently demonstrated promise in computer vision problems. However, unlike Convolutional Neural Networks (CNN), it is known that the performance of ViT saturates quickly with depth increasing, due to the observed attention collapse or patch uniformity. Despite a couple of empirical solutions, a rigorous framework studying on this scalability issue remains elusive. In this paper, we first establish a rigorous theory framework to analyze ViT features from the Fourier spectrum domain. We show that the self-attention mechanism inherently amounts to a low-pass filter, which indicates when ViT scales up its depth, excessive low-pass filtering will cause feature maps to only preserve their Direct-Current (DC) component. We then propose two straightforward yet effective techniques to mitigate the undesirable low-pass limitation. The first technique, termed AttnScale, decomposes a self-attention block into low-pass and high-pass components, then rescales and combines these two filters to produce an all-pass self-attention matrix. The second technique, termed FeatScale, re-weights feature maps on separate frequency bands to amplify the high-frequency signals. Both techniques are efficient and hyperparameter-free, while effectively overcoming relevant ViT training artifacts such as attention collapse and patch uniformity. By seamlessly plugging in our techniques to multiple ViT variants, we demonstrate that they consistently help ViTs benefit from deeper architectures, bringing up to 1.1% performance gains "for free" (e.g., with little parameter overhead). We publicly release our codes and pre-trained models at https://github.com/VITA-Group/ViT-Anti-Oversmoothing.

Self-ensemble Adversarial Training for Improved Robustness

- Hongjun Wang, Yisen Wang
- abstract@open-review(Poster): Due to numerous breakthroughs in real-world applications brought by machine intelligence, deep neural networks (DNNs) are widely employed in critical applications. However, predictions of DNNs are easily manipulated with imperceptible adversarial perturbations, which impedes the further deployment of DNNs and may result in profound security and privacy implications. By incorporating adversarial samples into the training data pool, adversarial training is the strongest principled strategy against various adversarial attacks among all sorts of defense methods. Recent works mainly focus on developing new loss functions or regularizers, attempting to find the unique optimal point in the weight space. But none of them taps the potentials of classifiers obtained from standard adversarial training, especially states on the searching trajectory of training. In this work, we are dedicated to the weight states of models through the training process and devise a simple but powerful \emph{Self-Ensemble Adversarial Training} (SEAT) method for yielding a robust classifier by averaging weights of history models. This considerably improves the robustness of the target model against several well known adversarial attacks, even merely utilizing the naive cross-entropy loss to supervise. We also discuss the relationship between

the ensemble of predictions from different adversarially trained models and the prediction of weight-ensembled models, as well as provide theoretical and empirical evidence that the proposed self-ensemble method provides a smoother loss landscape and better robustness than both individual models and the ensemble of predictions from different classifiers. We further analyze a subtle but fatal issue in the general settings for the self-ensemble model, which causes the deterioration of the weight-ensembled method in the late phases.

Do deep networks transfer invariances across classes?

- Allan Zhou, Fahim Tajwar, Alexander Robey, Tom Knowles, George J. Pappas, Hamed Hassani, Chelsea Finn
- abstract@open-review(Poster): In order to generalize well, classifiers must learn to be invariant to nuisance transformations that do not alter an input's class. Many problems have "class-agnostic" nuisance transformations that apply similarly to all classes, such as lighting and background changes for image classification. Neural networks can learn these invariances given sufficient data, but many real-world datasets are heavily class imbalanced and contain only a few examples for most of the classes. We therefore pose the question: how well do neural networks transfer class-agnostic invariances learned from the large classes to the small ones? Through careful experimentation, we observe that invariance to class-agnostic transformations is still heavily dependent on class size, with the networks being much less invariant on smaller classes. This result holds even when using data balancing techniques, and suggests poor invariance transfer across classes. Our results provide one explanation for why classifiers generalize poorly on unbalanced and long-tailed distributions. Based on this analysis, we show how a generative approach for learning the nuisance transformations can help transfer invariances across classes and improve performance on a set of imbalanced image classification benchmarks.

Cross-Trajectory Representation Learning for Zero-Shot Generalization in RL

- Bogdan Mazoure, Ahmed M Ahmed, R Devon Hjelm, Andrey Kolobov, Patrick MacAlpine
- abstract@open-review(Poster): A highly desirable property of a reinforcement learning (RL) agent -- and a major difficulty for deep RL approaches -- is the ability to generalize policies learned on a few tasks over a high-dimensional observation space to similar tasks not seen during training. Many promising approaches to this challenge consider RL as a process of training two functions simultaneously: a complex nonlinear encoder that maps high-dimensional observations to a latent representation space, and a simple linear policy over this space. We posit that a superior encoder for zero-shot generalization in RL can be trained by using solely an auxiliary SSL objective if the training process encourages the encoder to map behaviorally similar observations to similar representations, as reward-based signal can cause overfitting in the encoder (Raileanu et al., 2021). We propose Cross-Trajectory Representation Learning (CTRL), a method that runs within an RL agent and conditions its encoder to recognize behavioral similarity in observations by applying a novel SSL objective to pairs of trajectories from the agent's policies. CTRL can be viewed as having the same effect as inducing a pseudo-bisimulation metric but, crucially, avoids the use of rewards and associated overfitting risks. Our experiments ablate various components of CTRL and demonstrate that in combination with PPO it achieves better generalization performance on the challenging Procgen benchmark suite (Cobbe et al., 2020).

On Covariate Shift of Latent Confounders in Imitation and Reinforcement Learning

- Guy Tennenholtz, Assaf Hallak, Gal Dalal, Shie Mannor, Gal Chechik, Uri Shalit
- abstract@open-review(Poster): We consider the problem of using expert data with unobserved confounders for imitation and reinforcement learning. We begin by defining the problem of learning from confounded expert data in a contextual MDP setup. We analyze the limitations of learning from such data with and without external reward and propose an adjustment of standard imitation learning algorithms to fit this setup. In addition, we discuss the problem of distribution shift between the expert data and the online environment when partial observability is present in the data. We prove possibility and impossibility results for imitation learning under arbitrary distribution shift of the missing covariates. When additional external reward is provided, we propose a sampling procedure that addresses the unknown shift and prove convergence to an optimal solution. Finally, we validate our claims empirically on challenging assistive healthcare and recommender system simulation tasks.

RvS: What is Essential for Offline RL via Supervised Learning?

- Scott Emmons, Benjamin Eysenbach, Ilya Kostrikov, Sergey Levine
- abstract@open-review(Poster): Recent work has shown that supervised learning alone, without temporal difference (TD) learning, can be remarkably effective for offline RL. When does this hold true, and which algorithmic components are necessary? Through extensive experiments, we boil supervised learning for offline RL down to its essential elements. In every environment suite we consider, simply maximizing likelihood with a two-layer feedforward MLP is competitive with state-of-the-art results of substantially more complex methods based on TD learning or sequence modeling with Transformers. Carefully choosing model capacity (e.g., via regularization or architecture) and choosing which information to condition on (e.g., goals or rewards) are critical for performance. These insights serve as a field guide for practitioners doing Reinforcement Learning via Supervised Learning (which we coin RvS learning). They also probe the limits of existing RvS methods, which are comparatively weak on random data, and suggest a number of open problems.

LEARNING GUARANTEES FOR GRAPH CONVOLUTIONAL NETWORKS ON THE STOCHASTIC BLOCK MODEL

- Wei Lu
- abstract@open-review(Poster): An abundance of neural network models and algorithms for diverse tasks on graphs have been developed in the past five years. However, very few provable guarantees have been available for the performance of graph neural network models. This state of affairs is in contrast with the steady progress on the theoretical underpinnings of traditional dense and convolutional neural networks. In this paper we present the first provable guarantees for one of the best-studied families of graph neural network models, Graph Convolutional Networks (GCNs), for semi- supervised community detection tasks. We show that with high probability over the initialization and training data, a GCN will efficiently learn to detect communities on graphs drawn from a stochastic block model. Our proof relies on a fine-grained analysis of the training dynamics in order to overcome the complexity of a non-convex optimization landscape with many poorly-performing local minima.

Learning Versatile Neural Architectures by Propagating Network Codes

- Mingyu Ding, Yuqi Huo, Haoyu Lu, Linjie Yang, Zhe Wang, Zhiwu Lu, Jingdong Wang, Ping Luo
- abstract@open-review(Poster): This work explores how to design a single neural network capable of adapting to multiple heterogeneous vision tasks, such as image segmentation, 3D detection, and video recognition. This goal is challenging because both network architecture search (NAS) spaces and methods in different tasks are inconsistent. We solve this challenge from both sides. We first introduce a unified design space for multiple tasks and build a multitask NAS benchmark (NAS-Bench-MR) on many widely used datasets, including ImageNet, Cityscapes, KITTI, and HMDB51. We further propose Network Coding Propagation (NCP), which back-propagates gradients of neural predictors to directly update architecture codes along the desired gradient directions to solve various tasks. In this way, optimal architecture configurations can be found by NCP in our large search space in seconds.

Unlike prior arts of NAS that typically focus on a single task, NCP has several unique benefits. (1) NCP transforms architecture optimization from data-driven to architecture-driven, enabling joint search an architecture among multitasks with different data distributions. (2) NCP learns from network codes but not original data, enabling it to update the architecture efficiently across datasets. (3) In addition to our NAS-Bench-MR, NCP performs well on other NAS benchmarks, such as NAS-Bench-201. (4) Thorough studies of NCP on inter-, cross-, and intra-tasks highlight the importance of cross-task neural architecture design, i.e., multitask neural architectures and architecture transferring between different tasks. Code is available at https://github.com/dingmyu/NCP.

Task-Induced Representation Learning

- Jun Yamada, Karl Pertsch, Anisha Gunjal, Joseph J Lim
- abstract@open-review(Poster): In this work, we evaluate the effectiveness of representation learning approaches for decision making in visually complex environments. Representation learning is essential for effective reinforcement learning (RL) from high-dimensional in- puts. Unsupervised representation learning approaches based on reconstruction, prediction or contrastive learning have shown substantial learning efficiency gains. Yet, they have mostly been evaluated in clean laboratory or simulated settings. In contrast, real environments are visually complex and contain substantial amounts of clutter and distractors. Unsupervised representations will learn to model such distractors, potentially impairing the agentâ€™s learning efficiency. In contrast, an alternative class of approaches, which we call task-induced representation learning, leverages task information such as rewards or demonstrations from prior tasks to focus on task-relevant parts of the scene and ignore distractors. We investi- gate the effectiveness of unsupervised and task-induced representation learning approaches on four visually complex environments, from Distracting DMControl to the CARLA driving simulator. For both, RL and imitation learning, we find that representation learning generally improves sample efficiency on unseen tasks even in visually complex scenes and that task-induced representations can double learning efficiency compared to unsupervised alternatives.

Graph-based Nearest Neighbor Search in Hyperbolic Spaces

- Liudmila Prokhorenkova, Dmitry Baranchuk, Nikolay Bogachev, Yury Demidovich, Alexander Kolpakov
- abstract@open-review(Poster): The nearest neighbor search (NNS) problem is widely studied in Euclidean space, and graph-based algorithms are known to outperform other approaches for this task. However, hyperbolic geometry often allows for better data representation in various domains, including graphs, words, and images. In this paper, we show that graph-based approaches are also well suited for hyperbolic geometry. From a theoretical perspective, we rigorously analyze the time and space complexity of graph-based NNS, assuming that an n-element dataset is uniformly distributed within a d-dimensional ball of radius R in the hyperbolic space of curvature -1. Under some conditions on R and d, we derive the time and space complexity of graph-based NNS and compare the obtained results with known guarantees for the Euclidean case. Interestingly, in the dense setting ($d \ll \log n$) and under some assumptions on the radius R, graph-based NNS has lower time complexity in the hyperbolic space. This agrees with our experiments: we consider datasets embedded in hyperbolic and Euclidean spaces and show that graph-based NNS can be more efficient in the hyperbolic space. We also demonstrate that graph-based methods outperform other existing baselines for hyperbolic NNS. Overall, our theoretical and empirical analysis suggests that graph-based NNS can be considered a default approach for similarity search in hyperbolic spaces.

Generative Models as a Data Source for Multiview Representation Learning

- Ali Jahanian, Xavier Puig, Yonglong Tian, Phillip Isola
- abstract@open-review(Poster): Generative models are now capable of producing highly realistic images that look nearly indistinguishable from the data on which they are trained. This raises the question: if we have good enough generative models, do we still need datasets? We investigate this question in the setting of learning general-purpose visual representations from a black-box generative model rather than directly from data. Given an off-the-shelf image generator without any access to its training data, we train representations from the samples output by this generator. We compare several representation learning methods that can be applied to this setting, using the latent space of the generator to generate multiple "views" of the same semantic content. We show that for contrastive methods, this multiview data can naturally be used to identify positive pairs (nearby in latent space) and negative pairs (far apart in latent space). We find that the resulting representations rival or even outperform those learned directly from real data, but that good performance requires care in the sampling strategy applied and the training method. Generative models can be viewed as a compressed and organized copy of a dataset, and we envision a future where more and more "model zoos" proliferate while datasets become increasingly unwieldy, missing, or private. This paper suggests several techniques for dealing with visual representation learning in such a future. Code is available on our project page https://ali-design.github.io/GenRep/.

GiraffeDet: A Heavy-Neck Paradigm for Object Detection

- yiqi jiang, Zhiyu Tan, Junyan Wang, Xiuyu Sun, Ming Lin, Hao Li
- abstract@open-review(Poster): In conventional object detection frameworks, a backbone body inherited from image recognition models extracts deep latent features and then a neck module fuses these latent features to capture information at different scales. As the resolution in object detection is much larger than in image recognition, the computational cost of the backbone often dominates the total inference cost. This heavy-backbone design paradigm is mostly due to the historical legacy when transferring image recognition models to object detection rather than an end-to-end optimized design for object detection. In this work, we show that such paradigm indeed leads to sub-optimal object detection models. To this end, we propose a novel heavy-neck paradigm, GiraffeDet, a giraffe-like network for efficient object detection. The GiraffeDet uses an extremely lightweight backbone and a very deep and large neck module which encourages dense information exchange among different spatial scales as well as different levels of latent semantics simultaneously. This design paradigm allows detectors to process the high-level semantic information and low-level spatial information at the same priority even in the early stage of the network, making it more effective in detection tasks. Numerical evaluations on multiple popular object detection benchmarks show that GiraffeDet consistently outperforms previous SOTA models across a wide spectrum of resource constraints. The source code is available at https://github.com/jyqi/GiraffeDet.

A Unified Wasserstein Distributional Robustness Framework for Adversarial Training

- Anh Tuan Bui, Trung Le, Quan Hung Tran, He Zhao, Dinh Phung
- abstract@open-review(Poster): It is well-known that deep neural networks (DNNs) are susceptible to adversarial attacks, exposing a severe fragility of deep learning systems. As the result, adversarial training (AT) method, by incorporating adversarial examples during training, represents a natural and effective approach to strengthen the robustness of a DNN-based classifier. However, most AT-based methods, notably PGD-AT and TRADES, typically seek a pointwise adversary that generates the worst-case adversarial example by independently perturbing each data sample, as a way to ``probe'' the vulnerability of the classifier. Arguably, there are unexplored benefits in considering such adversarial effects from an entire distribution. To this end, this paper presents a unified framework that connects Wasserstein distributional robustness with current state-of-the-art AT methods. We introduce a new Wasserstein cost function and a new series of risk functions, with which we show that standard AT methods are special cases of their counterparts in our framework. This connection leads to an intuitive relaxation and generalization of existing AT methods and facilitates the development of a new family of distributional robustness AT-based algorithms. Extensive experiments show that our distributional robustness AT algorithms robustify further their standard AT counterparts in various settings.

miniF2F: a cross-system benchmark for formal Olympiad-level mathematics

- Kunhao Zheng, Jesse Michael Han, Stanislas Polu
- abstract@open-review(Poster): We present $\textsf{miniF2F}$, a dataset of formal Olympiad-level mathematics problems statements intended to provide a unified cross-system benchmark for neural theorem proving. The $\textsf{miniF2F}$ benchmark currently targets Metamath, Lean, Isabelle (partially) and HOL Light (partially) and consists of 488 problem statements drawn from the AIME, AMC, and the International Mathematical Olympiad (IMO), as well as material from high-school and undergraduate mathematics courses. We report baseline results using GPT-f, a neural theorem prover based on GPT-3 and provide an analysis of its performance. We intend for $\textsf{miniF2F}$ to be a community-driven effort and hope that our benchmark will help spur advances in neural theorem proving.

Towards Model Agnostic Federated Learning Using Knowledge Distillation

- Andrei Afonin, Sai Praneeth Karimireddy
- abstract@open-review(Poster): Is it possible to design an universal API for federated learning using which an ad-hoc group of data-holders (agents) collaborate with each other and perform federated learning? Such an API would necessarily need to be model-agnostic i.e. make no assumption about the model architecture being used by the agents, and also cannot rely on having representative public data at hand. Knowledge distillation (KD) is the obvious tool of choice to design such protocols. However, surprisingly, we show that most natural KD-based federated learning protocols have poor performance.

  To investigate this, we propose a new theoretical framework, Federated Kernel ridge regression, which can capture both model heterogeneity as well as data heterogeneity. Our analysis shows that the degradation is largely due to a fundamental limitation of knowledge distillation under data heterogeneity. We further validate our framework by analyzing and designing new protocols based on KD. Their performance on real world experiments using neural networks, though still unsatisfactory, closely matches our theoretical predictions.

Acceleration of Federated Learning with Alleviated Forgetting in Local Training

- Chencheng Xu, Zhiwei Hong, Minlie Huang, Tao Jiang
- abstract@open-review(Poster): Federated learning (FL) enables distributed optimization of machine learning models while protecting privacy by independently training local models on each client and then aggregating parameters on a central server, thereby producing an effective global model. Although a variety of FL algorithms have been proposed, their training efficiency remains low when the data are not independently and identically distributed (non-i.i.d.) across different clients. We observe that the slow convergence rates of the existing methods are (at least partially) caused by the catastrophic forgetting issue during the local training stage on each individual client, which leads to a large increase in the loss function concerning the previous training data provided at other clients. Here, we propose FedReg, an algorithm to accelerate FL with alleviated knowledge forgetting in the local training stage by regularizing locally trained parameters with the loss on generated pseudo data, which encode the knowledge of previous training data learned by the global model. Our comprehensive experiments demonstrate that FedReg not only significantly improves the convergence rate of FL, especially when the neural network architecture is deep and the clients' data are extremely non-i.i.d., but is also able to protect privacy better in classification problems and more robust against gradient inversion attacks.

Discovering Invariant Rationales for Graph Neural Networks

- Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, Tat-Seng Chua
- abstract@open-review(Poster): Intrinsic interpretability of graph neural networks (GNNs) is to find a small subset of the input graph's features --- rationale --- which guides the model prediction. Unfortunately, the leading rationalization models often rely on data biases, especially shortcut features, to compose rationales and make predictions without probing the critical and causal patterns. Moreover, such data biases easily change outside the training distribution. As a result, these models suffer from a huge drop in interpretability and predictive performance on out-of-distribution data. In this work, we propose a new strategy of discovering invariant rationale (DIR) to construct intrinsically interpretable GNNs. It conducts interventions on the training distribution to create multiple interventional distributions. Then it approaches the causal rationales that are invariant across different distributions while filtering out the spurious patterns that are unstable. Experiments on both synthetic and real-world datasets validate the superiority of our DIR in terms of interpretability and generalization ability on graph classification over the leading baselines. Code and datasets are available at https://github.com/Wuyxin/DIR-GNN.

Representing Mixtures of Word Embeddings with Mixtures of Topic Embeddings

- dongsheng wang, Dan dan Guo, He Zhao, Huangjie Zheng, Korawat Tanwisuth, Bo Chen, Mingyuan Zhou
- abstract@open-review(Poster): A topic model is often formulated as a generative model that explains how each word of a document is generated given a set of topics and document-specific topic proportions. It is focused on capturing the word co-occurrences in a document and hence often suffers from poor performance in analyzing short documents. In addition, its parameter estimation often relies on approximate posterior inference that is either not scalable or suffering from large approximation error. This paper introduces a new topic-modeling framework where each document is viewed as a set of word embedding vectors and each topic is modeled as an embedding vector in the same embedding space. Embedding the words and topics in the same vector space, we define a method to measure the semantic difference between the embedding vectors of the words of a document and these of the topics, and optimize the topic embeddings to minimize the expected difference over all documents. Experiments on text analysis demonstrate that the proposed method, which is amenable to mini-batch stochastic gradient descent based optimization and hence scalable to big corpora, provides competitive performance in discovering more coherent and diverse topics and extracting better document representations.

Generative Modeling with Optimal Transport Maps

- Litu Rout, Alexander Korotin, Evgeny Burnaev
- abstract@open-review(Poster): With the discovery of Wasserstein GANs, Optimal Transport (OT) has become a powerful tool for large-scale generative modeling tasks. In these tasks, OT cost is typically used as the loss for training GANs. In contrast to this approach, we show that the OT map itself can be used as a generative model, providing comparable performance. Previous analogous approaches consider OT maps as generative models only in the latent spaces due to their poor performance in the original high-dimensional ambient space. In contrast, we apply OT maps directly in the ambient space, e.g., a space of high-dimensional images. First, we derive a min-max optimization algorithm to efficiently compute OT maps for the quadratic cost (Wasserstein-2 distance). Next, we extend the approach to the case when the input and output distributions are located in the spaces of different dimensions and derive error bounds for the computed OT map. We evaluate the algorithm on image generation and unpaired image restoration tasks. In particular, we consider denoising, colorization, and inpainting, where the optimality of the restoration map is a desired attribute, since the output (restored) image is expected to be close to the input (degraded) one.

Focus on the Common Good: Group Distributional Robustness Follows

- Vihari Piratla, Praneeth Netrapalli, Sunita Sarawagi
- abstract@open-review(Poster): We consider the problem of training a classification model with group annotated training data. Recent work has established that, if there is distribution shift across different groups, models trained using the standard empirical risk minimization (ERM) objective suffer from poor performance on minority groups and that group distributionally robust optimization (Group-DRO) objective is a better alternative. The starting point of this paper is the observation that though Group-DRO performs better than ERM on minority groups for some benchmark datasets, there are several other datasets where it performs much worse than ERM. Inspired by ideas from the closely related problem of domain generalization, this paper proposes a new and simple algorithm that explicitly encourages learning of features that are shared across various groups. The key insight behind our proposed algorithm is that while Group-DRO focuses on groups with worst regularized loss, focusing instead, on groups that enable better performance even on other groups, could lead to learning of shared/common features, thereby enhancing minority performance beyond what is achieved by Group-DRO. Empirically, we show that our proposed algorithm matches or achieves better performance compared to strong contemporary baselines including ERM and Group-DRO on standard benchmarks on both minority groups and across all groups. Theoretically, we show that the proposed algorithm is a descent method and finds first order stationary points of smooth nonconvex functions.

Omni-Scale CNNs: a simple and effective kernel size configuration for time series classification

- Wensi Tang, Guodong Long, Lu Liu, Tianyi Zhou, Michael Blumenstein, Jing Jiang
- abstract@open-review(Poster): The size of the receptive field has been one of the most important factors for One Dimensional Convolutional Neural Networks (1D-CNNs) on time series classification tasks. Large efforts have been taken to choose the appropriate receptive field size, for it has a huge influence on the performance and differs significantly for each dataset. In this paper, we propose an Omni-Scale block (OS-block) for 1D-CNNs, where the kernel sizes are set by a simple and universal rule. OS-block can efficiently cover the best size of the receptive field across different datasets. This set of kernel sizes consists of multiple prime numbers according to the length of the time series. We experimentally show 1D-CNNs built from OS-block can consistently achieve the state-of-the-art accuracy with a smaller model size on five time series benchmarks, including both univariate and multivariate data from multiple domains. Comprehensive analysis and ablation studies shed light on how our rule finds the best receptive field size and demonstrate the consistency of our OS-block for multiple 1D-CNN structures.

Ada-NETS: Face Clustering via Adaptive Neighbour Discovery in the Structure Space

- Yaohua Wang, Yaobin Zhang, Fangyi Zhang, Senzhang Wang, Ming Lin, YuQi Zhang, Xiuyu Sun
- abstract@open-review(Poster): Face clustering has attracted rising research interest recently to take advantage of massive amounts of face images on the web. State-of-the-art performance has been achieved by Graph Convolutional Networks (GCN) due to their powerful representation capacity. However, existing GCN-based methods build face graphs mainly according to kNN relations in the feature space, which may lead to a lot of noise edges connecting two faces of different classes. The face features will be polluted when messages pass along these noise edges, thus degrading the performance of GCNs. In this paper, a novel algorithm named Ada-NETS is proposed to cluster faces by constructing clean graphs for GCNs. In Ada-NETS, each face is transformed to a new structure space, obtaining robust features by considering face features of the neighbour images. Then, an adaptive neighbour discovery strategy is proposed to determine a proper number of edges connecting to each face image. It significantly reduces the noise edges while maintaining the good ones to build a graph with clean yet rich edges for GCNs to cluster faces. Experiments on multiple public clustering datasets show that Ada-NETS significantly outperforms current state-of-the-art methods, proving its superiority and generalization. Code is available at https://github.com/damo-cv/Ada-NETS.

Decoupled Adaptation for Cross-Domain Object Detection

- Junguang Jiang, Baixu Chen, Jianmin Wang, Mingsheng Long
- abstract@open-review(Poster): Cross-domain object detection is more challenging than object classification since multiple objects exist in an image and the location of each object is unknown in the unlabeled target domain. As a result, when we adapt features of different objects to enhance the transferability of the detector, the features of the foreground and the background are easy to be confused, which may hurt the discriminability of the detector. Besides, previous methods focused on category adaptation but ignored another important part for object detection, i.e., the adaptation on bounding box regression. To this end, we propose D-adapt, namely Decoupled Adaptation, to decouple the adversarial adaptation and the training of the detector. Besides, we fill the blank of regression domain adaptation in object detection by introducing a bounding box adaptor. Experiments show that \textit{D-adapt} achieves state-of-the-art results on four cross-domain object detection tasks and yields 17\% and 21\% relative improvement on benchmark datasets Clipart1k and Comic2k in particular.

Rethinking Network Design and Local Geometry in Point Cloud: A Simple Residual MLP Framework

- Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, Yun Fu
- abstract@open-review(Poster): Point cloud analysis is challenging due to irregularity and unordered data structure. To capture the 3D geometries, prior works mainly rely on exploring sophisticated local geometric extractors, using convolution, graph, or attention mechanisms. These methods, however, incur unfavorable latency during inference and the performance saturates over the past few years. In this paper, we present an ovel perspective on this task. We find detailed local geometrical informationprobably is not the key to point cloud analysis â€“ we introduce a pure residual MLP network, called PointMLP, which integrates no local geometrical extractors but still performs very competitively. Equipped with a proposed lightweight geometric-affine module to stabilize the training, PointMLP delivers the new state-of-the-art on multiple datasets. On the real-world ScanObjectNN dataset, our method even surpasses the prior best method by 3.3% accuracy. We emphasize PointMLP achieves this strong performance without any sophisticated operations, hence leading to a prominent inference speed. Compared to most recent CurveNet, PointMLP trains 2Ã— faster, tests 7Ã— faster, and is more accurate on ModelNet40 benchmark. We hope our PointMLP may help the community towards a better understanding of point cloud analysis. The code is available at https://github.com/ma-xu/pointMLP-pytorch.

New Insights on Reducing Abrupt Representation Change in Online Continual Learning

- Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, Eugene Belilovsky
- abstract@open-review(Poster): In the online continual learning paradigm, agents must learn from a changing distribution while respecting memory and compute constraints. Experience Replay (ER), where a small subset of past data is stored and replayed alongside new data, has emerged as a simple and effective learning strategy. In this work, we focus on the change in representations of observed data that arises when previously unobserved classes appear in the incoming data stream, and new classes must be distinguished from previous ones. We shed new light on this question by showing that applying ER causes the newly added classesâ€™ representations to overlap significantly with the previous classes, leading to highly disruptive parameter updates. Based on this empirical analysis, we propose a new method which mitigates this issue by shielding the learned representations from drastic adaptation to accommodate new classes. We show that using an asymmetric update rule pushes new classes to adapt to the older ones (rather than the reverse), which is more effective especially at task boundaries, where much of the forgetting typically occurs. Empirical results show significant gains over strong baselines on standard continual learning benchmarks.

Demystifying Batch Normalization in ReLU Networks: Equivalent Convex Optimization Models and Implicit Regularization

- Tolga Ergen, Arda Sahiner, Batu Ozturkler, John M. Pauly, Morteza Mardani, Mert Pilanci
- abstract@open-review(Poster): Batch Normalization (BN) is a commonly used technique to accelerate and stabilize training of deep neural networks. Despite its empirical success, a full theoretical understanding of BN is yet to be developed. In this work, we analyze BN through the lens of convex optimization. We introduce an analytic framework based on convex duality to obtain exact convex representations of weight-decay regularized ReLU networks with BN, which can be trained in polynomial-time. Our analyses also show that optimal layer weights can be obtained as simple closed-form formulas in the high-dimensional and/or overparameterized regimes. Furthermore, we find that Gradient Descent provides an algorithmic bias effect on the standard non-convex BN network, and we design an approach to explicitly encode this implicit regularization into the convex objective. Experiments with CIFAR image classification highlight the effectiveness of this explicit regularization for mimicking and substantially improving the performance of standard BN networks.

Associated Learning: an Alternative to End-to-End Backpropagation that Works on CNN, RNN, and Transformer

- Dennis Y.H. Wu, Dinan Lin, Vincent Chen, Hung-Hsuan Chen
- abstract@open-review(Poster): This paper studies Associate Learning (AL), an alternative methodology to the end-to-end backpropagation (BP). We introduce the workflow to convert a neural network into a proper structure such that AL can be used to learn the weights for various types of neural networks. We compared AL and BP on some of the most successful types of neural networks -- Convolutional Neural Network (CNN), Recurrent Neural

Network (RNN), and Transformer. Experimental results show that AL consistently outperforms BP on various open datasets. We discuss possible reasons for AL's success and its limitations.

MetaShift: A Dataset of Datasets for Evaluating Contextual Distribution Shifts and Training Conflicts

- Weixin Liang, James Zou
- abstract@open-review(Poster): Understanding the performance of machine learning models across diverse data distributions is critically important for reliable applications. Motivated by this, there is a growing focus on curating benchmark datasets that capture distribution shifts. While valuable, the existing benchmarks are limited in that many of them only contain a small number of shifts and they lack systematic annotation about what is different across different shifts. We present MetaShiftâ€”a collection of 12,868 sets of natural images across 410 classesâ€”to address this challenge. We leverage the natural heterogeneity of Visual Genome and its annotations to construct MetaShift. The key construction idea is to cluster images using its metadata, which provides context for each image (e.g. â€œcats with carsâ€ or â€œcats in bathroomâ€) that represent distinct data distributions. MetaShift has two important benefits: first, it contains orders of magnitude more natural data shifts than previously available. Second, it provides explicit explanations of what is unique about each of its data sets and a distance score that measures the amount of distribution shift between any two of its data sets. We demonstrate the utility of MetaShift in benchmarking several recent proposals for training models to be robust to data shifts. We find that the simple empirical risk minimization performs the best when shifts are moderate and no method had a systematic advantage for large shifts. We also show how MetaShift can help to visualize conflicts between data subsets during model training.

FP-DETR: Detection Transformer Advanced by Fully Pre-training

- Wen Wang, Yang Cao, Jing Zhang, Dacheng Tao
- abstract@open-review(Poster): Large-scale pre-training has proven to be effective for visual representation learning on downstream tasks, especially for improving robustness and generalization. However, the recently developed detection transformers only employ pre-training on its backbone while leaving the key component, i.e., a 12-layer transformer, being trained from scratch, which prevents the model from above benefits. This separated training paradigm is mainly caused by the discrepancy between the upstream and downstream tasks. To mitigate the issue, we propose FP-DETR, a new method that Fully Pre-Trains an encoder-only transformer and smoothly fine-tunes it for object detection via a task adapter. Inspired by the success of textual prompts in NLP, we treat query positional embeddings as visual prompts to help the model attend to the target area (prompting) and recognize the object. To this end, we propose the task adapter which leverages self-attention to model the contextual relation between object query embedding. Experiments on the challenging COCO dataset demonstrate that our FP-DETR achieves competitive performance. Moreover, it enjoys better robustness to common corruptions and generalization to small-size datasets than state-of-the-art detection transformers. Code will be made publicly available at $\url{https://github.com/encounter1997/FP-DETR}$.

Efficient and Differentiable Conformal Prediction with General Function Classes

- Yu Bai, Song Mei, Huan Wang, Yingbo Zhou, Caiming Xiong
- abstract@open-review(Poster): Quantifying the data uncertainty in learning tasks is often done by learning a prediction interval or prediction set of the label given the input. Two commonly desired properties for learned prediction sets are \emph{valid coverage} and \emph{good efficiency} (such as low length or low cardinality). Conformal prediction is a powerful technique for learning prediction sets with valid coverage, yet by default its conformalization step only learns a single parameter, and does not optimize the efficiency over more expressive function classes. In this paper, we propose a generalization of conformal prediction to multiple learnable parameters, by considering the constrained empirical risk minimization (ERM) problem of finding the most efficient prediction set subject to valid empirical coverage. This meta-algorithm generalizes existing conformal prediction algorithms, and we show that it achieves approximate valid population coverage and near-optimal efficiency within class, whenever the function class in the conformalization step is low-capacity in a certain sense. Next, this ERM problem is challenging to optimize as it involves a non-differentiable coverage constraint. We develop a gradient-based algorithm for it by approximating the original constrained ERM using differentiable surrogate losses and Lagrangians. Experiments show that our algorithm is able to learn valid prediction sets and improve the efficiency significantly over existing approaches in several applications such as prediction intervals with improved length, minimum-volume prediction sets for multi-output regression, and label prediction sets for image classification.

Safe Neurosymbolic Learning with Differentiable Symbolic Execution

- Chenxi Yang, Swarat Chaudhuri
- abstract@open-review(Poster): We study the problem of learning verifiably safe parameters for programs that use neural networks as well as symbolic, human-written code. Such neurosymbolic programs arise in many safety-critical domains. However, because they need not be differentiable, it is hard to learn their parameters using existing gradient-based approaches to safe learning. Our method, Differentiable Symbolic Execution (DSE), samples control flow paths in a program, symbolically constructs worst-case "safety loss" along these paths, and backpropagates the gradients of these losses through program operations using a generalization of the REINFORCE estimator. We evaluate the method on a mix of synthetic tasks and real-world benchmarks. Our experiments show that DSE significantly outperforms the state-of-the-art DiffAI method on these tasks.

SimVLM: Simple Visual Language Model Pretraining with Weak Supervision

- Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, Yuan Cao
- abstract@open-review(Poster): With recent progress in joint modeling of visual and textual representations, Vision-Language Pretraining (VLP) has achieved impressive performance on many multimodal downstream tasks. However, the requirement for expensive annotations including clean image captions and regional labels limits the scalability of existing approaches, and complicates the pretraining procedure with the introduction of multiple dataset-specific objectives. In this work, we relax these constraints and present a minimalist pretraining framework, named Simple Visual Language Model (SimVLM). Unlike prior work, SimVLM reduces the training complexity by exploiting large-scale weak supervision, and is trained end-to-end with a single prefix language modeling objective. Without utilizing extra data or task-specific customization, the resulting model significantly outperforms previous pretraining methods and achieves new state-of-the-art results on a wide range of discriminative and generative vision-language benchmarks, including VQA (+3.74% vqa-score), NLVR2 (+1.17% accuracy), SNLI-VE (+1.37% accuracy) and image captioning tasks (+10.1% average CIDEr score). Furthermore, we demonstrate that SimVLM acquires strong generalization and transfer ability, enabling zero-shot behavior including open-ended visual question answering and cross-modality transfer.

Bundle Networks: Fiber Bundles, Local Trivializations, and a Generative Approach to Exploring Many-to-one Maps

- Nico Courts, Henry Kvinge
- abstract@open-review(Poster): Many-to-one maps are ubiquitous in machine learning, from the image recognition model that assigns a multitude of distinct images to the concept of â€œcatâ€ to the time series forecasting model which assigns a range of distinct time-series to a single scalar regression value. While the primary use of such models is naturally to associate correct output to each input, in many problems it is also useful to be able to explore, understand, and sample from a model's fibers, which are the set of input values x such that $f(x) = y,$ for fixed y in the output space. In this paper we show that popular generative architectures are ill-suited to such tasks. Motivated by this, we introduce a novel generative architecture, Bundle Networks, based on the concept of a fiber bundle from (differential) topology. BundleNets exploit the idea of a local trivialization wherein a space

can be locally decomposed into a product space that cleanly encodes the many-to-one nature of the map. By enforcing this decomposition in BundleNets and by utilizing state-of-the-art invertible components, investigating a network's fibers becomes natural.

Privacy Implications of Shuffling

- Casey Meehan, Amrita Roy Chowdhury, Kamalika Chaudhuri, Somesh Jha
- abstract@open-review(Poster): \ldp deployments are vulnerable to inference attacks as an adversary can link the noisy responses to their identity and subsequently, auxiliary information using the \textit{order} of the data. An alternative model, shuffle \textsf{DP}, prevents this by shuffling the noisy responses uniformly at random. However, this limits the data learnability -- only symmetric functions (input order agnostic) can be learned. In this paper, we strike a balance and show that systematic shuffling of the noisy responses can thwart specific inference attacks while retaining some meaningful data learnability. To this end, we propose a novel privacy guarantee, \name-privacy, that captures the privacy of the order of a data sequence. \name-privacy allows tuning the granularity at which the ordinal information is maintained, which formalizes the degree the resistance to inference attacks trading it off with data learnability. Additionally, we propose a novel shuffling mechanism that can achieve \name-privacy and demonstrate the practicality of our mechanism via evaluation on real-world datasets.

On the role of population heterogeneity in emergent communication

- Mathieu Rita, Florian Strub, Jean-Bastien Grill, Olivier Pietquin, Emmanuel Dupoux
- abstract@open-review(Poster): Populations have often been perceived as a structuring component for language to emerge and evolve: the larger the population, the more systematic the language. While this observation is widespread in the sociolinguistic literature, it has not been reproduced in computer simulations with neural agents. In this paper, we thus aim to clarify this apparent contradiction. We explore emergent language properties by varying agent population size in the speaker-listener Lewis Game. After reproducing the experimental paradox, we challenge the simulation assumption that the agent community is homogeneous. We first investigate how speaker-listener asymmetry alters language structure to examine two potential diversity factors: training speed and network capacity. We find out that emergent language properties are only altered by the relative difference of factors between speaker and listener, and not by their absolute values. From then, we leverage this observation to control population heterogeneity without introducing confounding factors. We finally show that introducing such training speed heterogeneities naturally sort out the initial paradox: larger simulated communities start developing more systematic and structured languages.

Hindsight is 20/20: Leveraging Past Traversals to Aid 3D Perception

- Yurong You, Katie Z Luo, Xiangyu Chen, Junan Chen, Wei-Lun Chao, Wen Sun, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger
- abstract@open-review(Poster): Self-driving cars must detect vehicles, pedestrians, and other trafi¬ c participants accurately to operate safely. Small, far-away, or highly occluded objects are particularly challenging because there is limited information in the LiDAR point clouds for detecting them. To address this challenge, we leverage valuable information from the past: in particular, data collected in past traversals of the same scene. We posit that these past data, which are typically discarded, provide rich contextual information for disambiguating the above-mentioned challenging cases. To this end, we propose a novel end-to-end trainable Hindsight framework to extract this contextual information from past traversals and store it in an easy-to-query data structure, which can then be leveraged to aid future 3D object detection of the same scene. We show that this framework is compatible with most modern 3D detection architectures and can substantially improve their average precision on multiple autonomous driving datasets, most notably by more than 300% on the challenging cases. Our code is available at https://github.com/YurongYou/Hindsight.

Language-driven Semantic Segmentation

- Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, Rene Ranftl
- abstract@open-review(Poster): We present LSeg, a novel model for language-driven semantic image segmentation. LSeg uses a text encoder to compute embeddings of descriptive input labels (e.g., `grass'' or`building") together with a transformer-based image encoder that computes dense per-pixel embeddings of the input image. The image encoder is trained with a contrastive objective to align pixel embeddings to the text embedding of the corresponding semantic class. The text embeddings provide a flexible label representation in which semantically similar labels map to similar regions in the embedding space (e.g., `cat'' andfurry"). This allows LSeg to generalize to previously unseen categories at test time, without retraining or even requiring a single additional training sample. We demonstrate that our approach achieves highly competitive zero-shot performance compared to existing zero- and few-shot semantic segmentation methods, and even matches the accuracy of traditional segmentation algorithms when a fixed label set is provided. Code and demo are available at https://github.com/isl-org/lang-seg.

Image BERT Pre-training with Online Tokenizer

- Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, Tao Kong
- abstract@open-review(Poster): The success of language Transformers is primarily attributed to the pretext task of masked language modeling (MLM), where texts are first tokenized into semantically meaningful pieces. In this work, we study masked image modeling (MIM) and indicate the necessity and challenges of using a semantically meaningful visual tokenizer. We present a self-supervised framework iBOT that can perform masked prediction with an online tokenizer. Specifically, we perform self-distillation on masked patch tokens and take the teacher network as the online tokenizer, along with self-distillation on the class token to acquire visual semantics. The online tokenizer is jointly learnable with the MIM objective and dispenses with a multi-stage training pipeline where the tokenizer needs to be pre-trained beforehand. We show the prominence of iBOT by achieving an 82.3% linear probing accuracy and an 87.8% fine-tuning accuracy evaluated on ImageNet-1K. Beyond the state-of-the-art image classification results, we underline emerging local semantic patterns, which helps the models to obtain strong robustness against common corruptions and achieve leading results on dense downstream tasks, e.g., object detection, instance segmentation, and semantic segmentation.

Accelerated Policy Learning with Parallel Differentiable Simulation

- Jie Xu, Viktor Makoviychuk, Yashraj Narang, Fabio Ramos, Wojciech Matusik, Animesh Garg, Miles Macklin
- abstract@open-review(Poster): Deep reinforcement learning can generate complex control policies, but requires large amounts of training data to work effectively. Recent work has attempted to address this issue by leveraging differentiable simulators. However, inherent problems such as local minima and exploding/vanishing numerical gradients prevent these methods from being generally applied to control tasks with complex contact-rich dynamics, such as humanoid locomotion in classical RL benchmarks. In this work we present a high-performance differentiable simulator and a new policy learning algorithm (SHAC) that can effectively leverage simulation gradients, even in the presence of non-smoothness. Our learning algorithm alleviates problems with local minima through a smooth critic function, avoids vanishing/exploding gradients through a truncated learning window, and allows many physical environments to be run in parallel. We evaluate our method on classical RL control tasks, and show substantial improvements in sample efficiency and wall-clock time over state-of-the-art RL and differentiable simulation-based algorithms. In addition, we demonstrate the scalability of our method by applying it to the challenging high-dimensional problem of muscle-actuated locomotion with a large action space, achieving a greater than $17\times$ reduction in training time over the best-performing established RL algorithm. More visual results are provided at: https://short-horizon-actor-critic.github.io/.

Do We Need Anisotropic Graph Neural Networks?

- Shyam A. Tailor, Felix Opolka, Pietro Lio, Nicholas Donald Lane
- abstract@open-review(Poster): Common wisdom in the graph neural network (GNN) community dictates that anisotropic models---in which messages sent between nodes are a function of both the source and target node---are required to achieve state-of-the-art performance. Benchmarks to date have demonstrated that these models perform better than comparable isotropic models---where messages are a function of the source node only. In this work we provide empirical evidence challenging this narrative: we propose an isotropic GNN, which we call Efficient Graph Convolution (EGC), that consistently outperforms comparable anisotropic models, including the popular GAT or PNA architectures by using spatially-varying adaptive filters. In addition to raising important questions for the GNN community, our work has significant real-world implications for efficiency. EGC achieves higher model accuracy, with lower memory consumption and latency, along with characteristics suited to accelerator implementation, while being a drop-in replacement for existing architectures. As an isotropic model, it requires memory proportional to the number of vertices in the graph ($\mathcal{O}(V)$); in contrast, anisotropic models require memory proportional to the number of edges ($\mathcal{O}(E)$). We demonstrate that EGC outperforms existing approaches across 6 large and diverse benchmark datasets, and conclude by discussing questions that our work raise for the community going forward. Code and pretrained models for our experiments are provided at https://github.com/shyam196/egc.

Is High Variance Unavoidable in RL? A Case Study in Continuous Control

- Johan Bjorck, Carla P Gomes, Kilian Q Weinberger
- abstract@open-review(Poster): Reinforcement learning (RL) experiments have notoriously high variance, and minor details can have disproportionately large effects on measured outcomes. This is problematic for creating reproducible research and also serves as an obstacle when applying RL to sensitive real-world applications. In this paper, we investigate causes for this perceived instability. To allow for an in-depth analysis, we focus on a specifically popular setup with high variance -- continuous control from pixels with an actor-critic agent. In this setting, we demonstrate that poor outlier runs which completely fail to learn are an important source of variance, but that weight initialization and initial exploration are not at fault. We show that one cause for these outliers is unstable network parametrization which leads to saturating nonlinearities. We investigate several fixes to this issue and find that simply normalizing penultimate features is surprisingly effective. For sparse tasks, we also find that partially disabling clipped double Q-learning decreases variance. By combining fixes we significantly decrease variances, lowering the average standard deviation across 21 tasks by a factor >3 for a state-of-the-art agent. This demonstrates that the perceived variance is not necessarily inherent to RL. Instead, it may be addressed via simple modifications and we argue that developing low-variance agents is an important goal for the RL community.

Simple GNN Regularisation for 3D Molecular Property Prediction and Beyond

- Jonathan Godwin, Michael Schaarschmidt, Alexander L Gaunt, Alvaro Sanchez-Gonzalez, Yulia Rubanova, Petar VeliÄ koviÄ‡, James Kirkpatrick, Peter Battaglia
- abstract@open-review(Poster): In this paper we show that simple noisy regularisation can be an effective way to address oversmoothing. We first argue that regularisers ad-dressing oversmoothing should both penalise node latent similarity and encourage meaningful node representations. From this observation we derive â€œNoisy Nodesâ€ ,a simple technique in which we corrupt the input graph with noise, and add a noise correcting node-level loss. The diverse node level loss encourages latent node diversity, and the denoising objective encourages graph manifold learning. Our regulariser applies well-studied methods in simple, straightforward ways which allow even generic architectures to overcome oversmoothing and achieve state of the art results on quantum chemistry tasks such as QM9 and Open Catalyst, and improve results significantly on Open Graph Benchmark (OGB) datasets. Our results suggest Noisy Nodes can serve as a complementary building block in the GNN toolkit.

Should We Be Pre-training? An Argument for End-task Aware Training as an Alternative

- Lucio M. Dery, Paul Michel, Ameet Talwalkar, Graham Neubig
- abstract@open-review(Poster): In most settings of practical concern, machine learning practitioners know in advance what end-task they wish to boost with auxiliary tasks. However, widely used methods for leveraging auxiliary data like pre-training and its continued-pretraining variant are end-task agnostic: they rarely, if ever, exploit knowledge of the target task. We study replacing end-task agnostic continued training of pre-trained language models with end-task aware training of said models. We argue that for sufficiently important end-tasks, the benefits of leveraging auxiliary data in a task-aware fashion can justify forgoing the traditional approach of obtaining generic, end-task agnostic representations as with (continued) pre-training. On three different low-resource NLP tasks from two domains, we demonstrate that multi-tasking the end-task and auxiliary objectives results in significantly better downstream task performance than the widely-used task-agnostic continued pre-training paradigm of Gururangan et al. (2020). We next introduce an online meta-learning algorithm that learns a set of multi-task weights to better balance among our multiple auxiliary objectives, achieving further improvements on end-task performance and data efficiency.

Learning Super-Features for Image Retrieval

- Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, Yannis Kalantidis
- abstract@open-review(Poster): Methods that combine local and global features have recently shown excellent performance on multiple challenging deep image retrieval benchmarks, but their use of local features raises at least two issues. First, these local features simply boil down to the localized map activations of a neural network, and hence can be extremely redundant. Second, they are typically trained with a global loss that only acts on top of an aggregation of local features; by contrast, testing is based on local feature matching, which creates a discrepancy between training and testing. In this paper, we propose a novel architecture for deep image retrieval, based solely on mid-level features that we call Super-features. These Super-features are constructed by an iterative attention module and constitute an ordered set in which each element focuses on a localized and discriminant image pattern. For training, they require only image labels. A contrastive loss operates directly at the level of Super-features and focuses on those that match across images. A second complementary loss encourages diversity. Experiments on common landmark retrieval benchmarks validate that Super-features substantially outperform state-of-the-art methods when using the same number of features, and only require a significantly smaller memory footprint to match their performance. Code and models are available at: https://github.com/naver/FIRe.

Online Facility Location with Predictions

- Shaofeng H.-C. Jiang, Erzhi Liu, You Lyu, Zhihao Gavin Tang, Yubo Zhang
- abstract@open-review(Poster): We provide nearly optimal algorithms for online facility location (OFL) with predictions. In OFL, n demand points arrive in order and the algorithm must irrevocably assign each demand point to an open facility upon its arrival. The objective is to minimize the total connection costs from demand points to assigned facilities plus the facility opening cost. We further assume the algorithm is additionally given for each demand point x_i a natural prediction $f_{x_i}^{\mathrm{pred}}$ which is supposed to be the facility $f_{x_i}^{\mathrm{opt}}$ that serves x_i in the offline optimal solution.

Our main result is an $O(\min{\log {\frac{n\eta_\infty}{\mathrm{OPT}}}, \log{n} })$-competitive algorithm where η_∞ is the maximum prediction error (i.e., the distance between $f_{x_i}^{\mathrm{pred}}$ and $f_{x_i}^{\mathrm{opt}}$). Our algorithm overcomes the fundamental $\Omega(\frac{\log n}{\log \log n})$ lower bound of OFL (without predictions) when η_∞ is small, and it still maintains $O(\log n)$ ratio even when η_∞ is unbounded. Furthermore, our theoretical analysis is supported by empirical evaluations for the tradeoffs between η_∞ and the competitive ratio on various real datasets of different types.

Few-Shot Backdoor Attacks on Visual Object Tracking

- Yiming Li, Haoxiang Zhong, Xingjun Ma, Yong Jiang, Shu-Tao Xia
- abstract@open-review(Poster): Visual object tracking (VOT) has been widely adopted in mission-critical applications, such as autonomous driving and intelligent surveillance systems. In current practice, third-party resources such as datasets, backbone networks, and training platforms are frequently used to train high-performance VOT models. Whilst these resources bring certain convenience, they also introduce new security threats into VOT models. In this paper, we reveal such a threat where an adversary can easily implant hidden backdoors into VOT models by tempering with the training process. Specifically, we propose a simple yet effective few-shot backdoor attack (FSBA) that optimizes two losses alternately: 1) a \emph{feature loss} defined in the hidden feature space, and 2) the standard \emph{tracking loss}. We show that, once the backdoor is embedded into the target model by our FSBA, it can trick the model to lose track of specific objects even when the \emph{trigger} only appears in one or a few frames. We examine our attack in both digital and physical-world settings and show that it can significantly degrade the performance of state-of-the-art VOT trackers. We also show that our attack is resistant to potential defenses, highlighting the vulnerability of VOT models to potential backdoor attacks.

Backdoor Defense via Decoupling the Training Process

- Kunzhe Huang, Yiming Li, Baoyuan Wu, Zhan Qin, Kui Ren
- abstract@open-review(Poster): Recent studies have revealed that deep neural networks (DNNs) are vulnerable to backdoor attacks, where attackers embed hidden backdoors in the DNN model by poisoning a few training samples. The attacked model behaves normally on benign samples, whereas its prediction will be maliciously changed when the backdoor is activated. We reveal that poisoned samples tend to cluster together in the feature space of the attacked DNN model, which is mostly due to the end-to-end supervised training paradigm. Inspired by this observation, we propose a novel backdoor defense via decoupling the original end-to-end training process into three stages. Specifically, we first learn the backbone of a DNN model via \emph{self-supervised learning} based on training samples without their labels. The learned backbone will map samples with the same ground-truth label to similar locations in the feature space. Then, we freeze the parameters of the learned backbone and train the remaining fully connected layers via standard training with all (labeled) training samples. Lastly, to further alleviate side-effects of poisoned samples in the second stage, we remove labels of some `low-credible' samples determined based on the learned model and conduct a \emph{semi-supervised fine-tuning} of the whole model. Extensive experiments on multiple benchmark datasets and DNN models verify that the proposed defense is effective in reducing backdoor threats while preserving high accuracy in predicting benign samples. Our code is available at \url{https://github.com/SCLBD/DBD}.

Learning to Complete Code with Sketches

- Daya Guo, Alexey Svyatkovskiy, Jian Yin, Nan Duan, Marc Brockschmidt, Miltiadis Allamanis
- abstract@open-review(Poster): Code completion is usually cast as a language modelling problem, i.e., continuing an input in a left-to-right fashion. However, in practice, some parts of the completion (e.g., string literals) may be very hard to predict, whereas subsequent parts directly follow from the context. To handle this, we instead consider the scenario of generating code completions with "holes" inserted in places where a model is uncertain. We develop Grammformer, a Transformer-based model that guides the code generation by the programming language grammar, and compare it to a variety of more standard sequence models.

We train the models on code completion for C# and Python given partial code context. To evaluate models, we consider both ROUGE as well as a new metric RegexAcc that measures success of generating completions matching long outputs with as few holes as possible. In our experiments, Grammformer generates 10-50% more accurate completions compared to traditional generative models and 37-50% longer sketches compared to sketch-generating baselines trained with similar techniques.

Reverse Engineering of Imperceptible Adversarial Image Perturbations

- Yifan Gong, Yuguang Yao, Yize Li, Yimeng Zhang, Xiaoming Liu, Xue Lin, Sijia Liu
- abstract@open-review(Poster): It has been well recognized that neural network based image classifiers are easily fooled by images with tiny perturbations crafted by an adversary. There has been a vast volume of research to generate and defend such adversarial attacks. However, the following problem is left unexplored: How to reverse-engineer adversarial perturbations from an adversarial image? This leads to a new adversarial learning paradigmâ€”Reverse Engineering of Deceptions (RED). If successful, RED allows us to estimate adversarial perturbations and recover the original images. However, carefully crafted, tiny adversarial perturbations are difficult to recover by optimizing a unilateral RED objective. For example, the pure image denoising method may overfit to minimizing the reconstruction error but hardly preserve the classification properties of the true adversarial perturbations. To tackle this challenge, we formalize the RED problem and identify a set of principles crucial to the RED approach design. Particularly, we find that prediction alignment and proper data augmentation (in terms of spatial transformations) are two criteria to achieve a generalizable RED approach. By integrating these RED principles with image denoising, we propose a new Class-Discriminative Denoising based RED framework, termed CDD-RED. Extensive experiments demonstrate the effectiveness of CDD-RED under different evaluation metrics (ranging from the pixel-level, prediction-level to the attribution-level alignment) and a variety of attack generation methods (e.g., FGSM, PGD, CW, AutoAttack, and adaptive attacks).

DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR

- Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, Lei Zhang
- abstract@open-review(Poster): We present in this paper a novel query formulation using dynamic anchor boxes for DETR (DEtection TRansformer) and offer a deeper understanding of the role of queries in DETR. This new formulation directly uses box coordinates as queries in Transformer decoders and dynamically updates them layer by layer. Using box coordinates not only helps using explicit positional priors to improve the query-to-feature similarity and eliminate the slow training convergence issue in DETR, but also allows us to modulate the positional attention map using the box width and height information. Such a design makes it clear that queries in DETR can be implemented as performing soft ROI pooling layer by layer in a cascade manner. As a result, it leads to the best performance on the MS-COCO benchmark among the DETR-like detection models under the same setting, e.g., AP 45.7\% using ResNet50-DC5 as backbone trained in 50 epochs. We also conducted extensive experiments to confirm our analysis and verify the effectiveness of our methods. Code is available at \url{https://github.com/IDEA-opensource/DAB-DETR}.

On the Certified Robustness for Ensemble Models and Beyond

- Zhuolin Yang, Linyi Li, Xiaojun Xu, Bhavya Kailkhura, Tao Xie, Bo Li
- abstract@open-review(Poster): Recent studies show that deep neural networks (DNN) are vulnerable to adversarial examples, which aim to mislead DNNs by adding perturbations with small magnitude. To defend against such attacks, both empirical and theoretical defense approaches have been extensively studied for a single ML model. In this work, we aim to analyze and provide the certified robustness for ensemble ML models, together with the sufficient and necessary conditions of robustness for different ensemble protocols. Although ensemble models are shown more robust than a single model empirically; surprisingly, we find that in terms of the certified robustness the standard ensemble models only achieve marginal improvement compared to a single model. Thus, to explore the conditions that guarantee to provide certifiably robust ensemble ML models, we first prove that diversified gradient and large confidence margin are sufficient and necessary conditions for certifiably robust ensemble models under the model-smoothness assumption. We then provide the bounded model-smoothness analysis based on the proposed Ensemble-before-Smoothing strategy. We also prove that an ensemble model can always achieve higher certified robustness than a single base model under mild conditions. Inspired by the theoretical findings, we propose the lightweight Diversity Regularized Training (DRT) to train certifiably robust ensemble ML models. Extensive experiments show

that our DRT enhanced ensembles can consistently achieve higher certified robustness than existing single and ensemble ML models, demonstrating the state-of-the-art certified L_2-robustness on MNIST, CIFAR-10, and ImageNet datasets.

Efficient Neural Causal Discovery without Acyclicity Constraints

- Phillip Lippe, Taco Cohen, Efstratios Gavves
- abstract@open-review(Poster): Learning the structure of a causal graphical model using both observational and interventional data is a fundamental problem in many scientific fields. A promising direction is continuous optimization for score-based methods, which, however, require constrained optimization to enforce acyclicity or lack convergence guarantees. In this paper, we present ENCO, an efficient structure learning method for directed, acyclic causal graphs leveraging observational and interventional data. ENCO formulates the graph search as an optimization of independent edge likelihoods, with the edge orientation being modeled as a separate parameter. Consequently, we provide for ENCO convergence guarantees under mild conditions, without having to constrain the score function with respect to acyclicity. In experiments, we show that ENCO can efficiently recover graphs with hundreds of nodes, an order of magnitude larger than what was previously possible, while handling deterministic variables and discovering latent confounders.

Pseudo-Labeled Auto-Curriculum Learning for Semi-Supervised Keypoint Localization

- Can Wang, Sheng Jin, Yingda Guan, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang
- abstract@open-review(Poster): Localizing keypoints of an object is a basic visual problem. However, supervised learning of a keypoint localization network often requires a large amount of data, which is expensive and time-consuming to obtain. To remedy this, there is an ever-growing interest in semi-supervised learning (SSL), which leverages a small set of labeled data along with a large set of unlabeled data. Among these SSL approaches, pseudo-labeling (PL) is one of the most popular. PL approaches apply pseudo-labels to unlabeled data, and then train the model with a combination of the labeled and pseudo-labeled data iteratively. The key to the success of PL is the selection of high-quality pseudo-labeled samples. Previous works mostly select training samples by manually setting a single confidence threshold. We propose to automatically select reliable pseudo-labeled samples with a series of dynamic thresholds, which constitutes a learning curriculum.Extensive experiments on five keypoint localization benchmark datasets demonstrate that the proposed approach significantly outperforms the previous state-of-the-art SSL approaches.

Signing the Supermask: Keep, Hide, Invert

- Nils Koster, Oliver Grothe, Achim Rettinger
- abstract@open-review(Poster): The exponential growth in numbers of parameters of neural networks over the past years has been accompanied by an increase in performance across several fields. However, due to their sheer size, the networks not only became difficult to interpret but also problematic to train and use in real-world applications, since hardware requirements increased accordingly. Tackling both issues, we present a novel approach that either drops a neural network's initial weights or inverts their respective sign. Put simply, a network is trained by weight selection and inversion without changing their absolute values. Our contribution extends previous work on masking by additionally sign-inverting the initial weights and follows the findings of the Lottery Ticket Hypothesis. Through this extension and adaptations of initialization methods, we achieve a pruning rate of up to 99%, while still matching or exceeding the performance of various baseline and previous models. Our approach has two main advantages. First, and most notable, signed Supermask models drastically simplify a model's structure, while still performing well on given tasks. Second, by reducing the neural network to its very foundation, we gain insights into which weights matter for performance. The code is available on GitHub.

Bootstrapping Semantic Segmentation with Regional Contrast

- Shikun Liu, Shuaifeng Zhi, Edward Johns, Andrew Davison
- abstract@open-review(Poster): We present ReCo, a contrastive learning framework designed at a regional level to assist learning in semantic segmentation. ReCo performs pixel-level contrastive learning on a sparse set of hard negative pixels, with minimal additional memory footprint. ReCo is easy to implement, being built on top of off-the-shelf segmentation networks, and consistently improves performance, achieving more accurate segmentation boundaries and faster convergence. The strongest effect is in semi-supervised learning with very few labels. With ReCo, we achieve high quality semantic segmentation model, requiring only 5 examples of each semantic class.

Generative Principal Component Analysis

- Zhaoqiang Liu, Jiulong Liu, Subhroshekhar Ghosh, Jun Han, Jonathan Scarlett
- abstract@open-review(Poster): In this paper, we study the problem of principal component analysis with generative modeling assumptions, adopting a general model for the observed matrix that encompasses notable special cases, including spiked matrix recovery and phase retrieval. The key assumption is that the first principal eigenvector lies near the range of an L-Lipschitz continuous generative model with bounded k-dimensional inputs. We propose a quadratic estimator, and show that it enjoys a statistical rate of order $\sqrt{\frac{k\log L}{m}}$, where m is the number of samples. Moreover, we provide a variant of the classic power method, which projects the calculated data onto the range of the generative model during each iteration. We show that under suitable conditions, this method converges exponentially fast to a point achieving the above-mentioned statistical rate. This rate is conjectured in~\citep{aubin2019spiked,cocola2020nonasymptotic} to be the best possible even when we only restrict to the special case of spiked matrix models. We perform experiments on various image datasets for spiked matrix and phase retrieval models, and illustrate performance gains of our method to the classic power method and the truncated power method devised for sparse principal component analysis.

Pareto Policy Pool for Model-based Offline Reinforcement Learning

- Yijun Yang, Jing Jiang, Tianyi Zhou, Jie Ma, Yuhui Shi
- abstract@open-review(Poster): Online reinforcement learning (RL) can suffer from poor exploration, sparse reward, insufficient data, and overhead caused by inefficient interactions between an immature policy and a complicated environment. Model-based offline RL instead trains an environment model using a dataset of pre-collected experiences so online RL methods can learn in an offline manner by solely interacting with the model. However, the uncertainty and accuracy of the environment model can drastically vary across different state-action pairs so the RL agent may achieve high model return but perform poorly in the true environment. Unlike previous works that need to carefully tune the trade-off between the model return and uncertainty in a single objective, we study a bi-objective formulation for model-based offline RL that aims at producing a pool of diverse policies on the Pareto front performing different levels of trade-offs, which provides the flexibility to select the best policy for each realistic environment from the pool. Our method, "Pareto policy pool (P3)", does not need to tune the trade-off weight but can produce policies allocated at different regions of the Pareto front. For this purpose, we develop an efficient algorithm that solves multiple bi-objective optimization problems with distinct constraints defined by reference vectors targeting diverse regions of the Pareto front. We theoretically prove that our algorithm can converge to the targeted regions. In order to obtain more Pareto optimal policies without linearly increasing the cost, we leverage the achieved policies as initialization to find more Pareto optimal policies in their neighborhoods. On the D4RL benchmark for offline RL, P3 substantially outperforms several recent baseline methods over multiple tasks, especially when the quality of pre-collected experiences is low.

Filling the G_ap_s: Multivariate Time Series Imputation by Graph Neural Networks

- Andrea Cini, Ivan Marisca, Cesare Alippi
- abstract@open-review(Poster): Dealing with missing values and incomplete time series is a labor-intensive, tedious, inevitable task when handling data coming from real-world applications. Effective spatio-temporal representations would allow imputation methods to reconstruct missing temporal data by exploiting information coming from sensors at different locations. However, standard methods fall short in capturing the nonlinear time and space dependencies existing within networks of interconnected sensors and do not take full advantage of the available - and often strong - relational information. Notably, most state-of-the-art imputation methods based on deep learning do not explicitly model relational aspects and, in any case, do not exploit processing frameworks able to adequately represent structured spatio-temporal data. Conversely, graph neural networks have recently surged in popularity as both expressive and scalable tools for processing sequential data with relational inductive biases. In this work, we present the first assessment of graph neural networks in the context of multivariate time series imputation. In particular, we introduce a novel graph neural network architecture, named GRIN, which aims at reconstructing missing data in the different channels of a multivariate time series by learning spatio-temporal representations through message passing. Empirical results show that our model outperforms state-of-the-art methods in the imputation task on relevant real-world benchmarks with mean absolute error improvements often higher than 20%.

An Unconstrained Layer-Peeled Perspective on Neural Collapse

- Wenlong Ji, Yiping Lu, Yiliang Zhang, Zhun Deng, Weijie J Su
- abstract@open-review(Poster): Neural collapse is a highly symmetric geometry of neural networks that emerges during the terminal phase of training, with profound implications on the generalization performance and robustness of the trained networks. To understand how the last-layer features and classifiers exhibit this recently discovered implicit bias, in this paper, we introduce a surrogate model called the unconstrained layer-peeled model (ULPM). We prove that gradient flow on this model converges to critical points of a minimum-norm separation problem exhibiting neural collapse in its global minimizer. Moreover, we show that the ULPM with the cross-entropy loss has a benign global landscape for its loss function, which allows us to prove that all the critical points are strict saddle points except the global minimizers that exhibit the neural collapse phenomenon. Empirically, we show that our results also hold during the training of neural networks in real-world tasks when explicit regularization or weight decay is not used.

Contrastive Clustering to Mine Pseudo Parallel Data for Unsupervised Translation

- Xuan-Phi Nguyen, Hongyu Gong, Yun Tang, Changhan Wang, Philipp Koehn, Shafiq Joty
- abstract@open-review(Poster): Modern unsupervised machine translation systems mostly train their models by generating synthetic parallel training data from large unlabeled monolingual corpora of different languages through various means, such as iterative back-translation. However, there may exist small amount of actual parallel data hidden in the sea of unlabeled data, which has not been exploited. We develop a new fine-tuning objective, called Language-Agnostic Constraint for SwAV loss, or LAgSwAV, which enables a pre-trained model to extract such pseudo-parallel data from the monolingual corpora in a fully unsupervised manner. We then propose an effective strategy to utilize the obtained synthetic data to augment unsupervised machine translation. Our method achieves the state of the art in the WMT'14 English-French, WMT'16 German-English and English-Romanian bilingual unsupervised translation tasks, with 40.2, 36.8, 37.0 BLEU, respectively. We also achieve substantial improvements in the FLoRes low-resource English-Nepali and English-Sinhala unsupervised tasks with 5.3 and 5.4 BLEU, respectively.

Multimeasurement Generative Models

- Saeed Saremi, Rupesh Kumar Srivastava
- abstract@open-review(Poster): We formally map the problem of sampling from an unknown distribution with a density in $\mathbb{R}^d$ to the problem of learning and sampling a smoother density in $\mathbb{R}^{Md}$ obtained by convolution with a fixed factorial kernel: the new density is referred to as M-density and the kernel as multimeasurement noise model (MNM). The M-density in $\mathbb{R}^{Md}$ is smoother than the original density in $\mathbb{R}^d$, easier to learn and sample from, yet for large M the two problems are mathematically equivalent since clean data can be estimated exactly given a multimeasurement noisy observation using the Bayes estimator. To formulate the problem, we derive the Bayes estimator for Poisson and Gaussian MNMs in closed form in terms of the unnormalized M-density. This leads to a simple least-squares objective for learning parametric energy and score functions. We present various parametrization schemes of interest including one in which studying Gaussian M-densities directly leads to multidenoising autoencodersâ€”this is the first theoretical connection made between denoising autoencoders and empirical Bayes in the literature. Samples in $\mathbb{R}^d$ are obtained by walk-jump sampling (Saremi & Hyvarinen, 2019) via underdamped Langevin MCMC (walk) to sample from M-density and the multimeasurement Bayes estimation (jump). We study permutation invariant Gaussian M-densities on MNIST, CIFAR-10, and FFHQ-256 datasets, and demonstrate the effectiveness of this framework for realizing fast-mixing stable Markov chains in high dimensions.

Information Gain Propagation: a New Way to Graph Active Learning with Soft Labels

- Wentao Zhang, Yexin Wang, Zhenbang You, Meng Cao, Ping Huang, Jiulong Shan, Zhi Yang, Bin CUI
- abstract@open-review(Poster): Graph Neural Networks (GNNs) have achieved great success in various tasks, but their performance highly relies on a large number of labeled nodes, which typically requires considerable human effort. GNN-based Active Learning (AL) methods are proposed to improve the labeling efficiency by selecting the most valuable nodes to label. Existing methods assume an oracle can correctly categorize all the selected nodes and thus just focus on the node selection. However, such an exact labeling task is costly, especially when the categorization is out of the domain of individual expert (oracle). The paper goes further, presenting a soft-label approach to AL on GNNs. Our key innovations are: i) relaxed queries where a domain expert (oracle) only judges the correctness of the predicted labels (a binary question) rather than identifying the exact class (a multi-class question), and ii) new criteria of maximizing information gain propagation for active learner with relaxed queries and soft labels. Empirical studies on public datasets demonstrate that our method significantly outperforms the state-of-the-art GNN-based AL methods in terms of both accuracy and labeling cost.

Constructing Orthogonal Convolutions in an Explicit Manner

- Tan Yu, Jun Li, YUNFENG CAI, Ping Li
- abstract@open-review(Poster): Convolutions with orthogonal input-output Jacobian matrix, i.e., orthogonal convolution, have recently attracted substantial attention. A convolution layer with an orthogonal Jacobian matrix is 1-Lipschitz in the 2-norm, making the output robust to the perturbation in input. Meanwhile, an orthogonal Jacobian matrix preserves the gradient norm in back-propagation, which is critical for stable training deep networks. Nevertheless, existing orthogonal convolutions are burdened by high computational costs for preserving orthogonality. In this work, we exploit the relation between the singular values of the convolution layer's Jacobian and the structure of the convolution kernel. To achieve orthogonality, we explicitly construct the convolution kernel for enforcing all singular values of the convolution layer's Jacobian to be 1s. After training, the explicitly constructed orthogonal (ECO) convolution is constructed only once, and their weights are stored. Then, in evaluation, we only need to load the stored weights of the trained ECO convolution, and the computational cost of ECO convolution is the same as the standard dilated convolution. It is more efficient than the recent state-of-the-art approach, skew orthogonal convolution (SOC) in evaluation. Experiments on CIFAR-10 and CIFAR-100 demonstrate that the proposed ECO convolution is faster than SOC in evaluation while leading to competitive standard and certified robust accuracies.

X-model: Improving Data Efficiency in Deep Learning with A Minimax Model

- Ximei Wang, Xinyang Chen, Jianmin Wang, Mingsheng Long
- abstract@open-review(Poster): To mitigate the burden of data labeling, we aim at improving data efficiency for both classification and regression setups in deep learning. However, the current focus is on classification problems while rare attention has been paid to deep regression, which usually requires more

human effort to labeling. Further, due to the intrinsic difference between categorical and continuous label space, the common intuitions for classification, \textit{e.g.} cluster assumptions or pseudo labeling strategies, cannot be naturally adapted into deep regression. To this end, we first delved into the existing data-efficient methods in deep learning and found that they either encourage invariance to \textit{data stochasticity} (\textit{e.g.}, consistency regularization under different augmentations) or \textit{model stochasticity} (\textit{e.g.}, difference penalty for predictions of models with different dropout). To take the power of both worlds, we propose a novel \Chi-model by simultaneously encouraging the invariance to {data stochasticity} and {model stochasticity}. Further, the \Chi-model plays a minimax game between the feature extractor and task-specific heads to further enhance the invariance to model stochasticity. Extensive experiments verify the superiority of the \Chi-model among various tasks, from a single-value prediction task of age estimation to a dense-value prediction task of keypoint localization, a 2D synthetic and a 3D realistic dataset, as well as a multi-category object recognition task.

Stein Latent Optimization for Generative Adversarial Networks

- Uiwon Hwang, Heeseung Kim, Dahuin Jung, Hyemi Jang, Hyungyu Lee, Sungroh Yoon
- abstract@open-review(Poster): Generative adversarial networks (GANs) with clustered latent spaces can perform conditional generation in a completely unsupervised manner. In the real world, the salient attributes of unlabeled data can be imbalanced. However, most of existing unsupervised conditional GANs cannot cluster attributes of these data in their latent spaces properly because they assume uniform distributions of the attributes. To address this problem, we theoretically derive Stein latent optimization that provides reparameterizable gradient estimations of the latent distribution parameters assuming a Gaussian mixture prior in a continuous latent space. Structurally, we introduce an encoder network and novel unsupervised conditional contrastive loss to ensure that data generated from a single mixture component represent a single attribute. We confirm that the proposed method, named Stein Latent Optimization for GANs (SLOGAN), successfully learns balanced or imbalanced attributes and achieves state-of-the-art unsupervised conditional generation performance even in the absence of attribute information (e.g., the imbalance ratio). Moreover, we demonstrate that the attributes to be learned can be manipulated using a small amount of probe data.

Sparse DETR: Efficient End-to-End Object Detection with Learnable Sparsity

- Byungseok Roh, JaeWoong Shin, Wuhyun Shin, Saehoon Kim
- abstract@open-review(Poster): DETR is the first end-to-end object detector using a transformer encoder-decoder architecture and demonstrates competitive performance but low computational efficiency. The subsequent work, Deformable DETR, enhances the efficiency of DETR by replacing dense attention with deformable attention, which achieves 10x faster convergence and improved performance. Using the multiscale feature to ameliorate performance, however, the number of encoder queries increases by 20x compared to DETR, and the computation cost of the encoder attention remains a bottleneck. We observe that the encoder queries referenced by the decoder account for only 45% of the total, and find out the detection accuracy does not deteriorate significantly even if only the referenced queries are polished in the encoder block. Inspired by this observation, we propose Sparse DETR that selectively updates only the queries expected to be referenced by the decoder, thus help the model effectively detect objects. In addition, we show that applying an auxiliary detection loss on the selected queries in the encoder improves the performance while minimizing computational overhead. We validate that Sparse DETR achieves better performance than Deformable DETR even with only 10% encoder queries on the COCO dataset. Albeit only the encoder queries are sparsified, the total computation cost decreases by 38% and the frames per second (FPS) increases by 42% compared to Deformable DETR. Code will be released.

Online Target Q-learning with Reverse Experience Replay: Efficiently finding the Optimal Policy for Linear MDPs

- Naman Agarwal, Syomantak Chaudhuri, Prateek Jain, Dheeraj Mysore Nagaraj, Praneeth Netrapalli
- abstract@open-review(Poster): Q-learning is a popular Reinforcement Learning (RL) algorithm which is widely used in practice with function approximation (Mnih et al., 2015). In contrast, existing theoretical results are pessimistic about Q-learning. For example, (Baird, 1995) shows that Q-learning does not converge even with linear function approximation for linear MDPs. Furthermore, even for tabular MDPs with synchronous updates, Q-learning was shown to have sub-optimal sample complexity (Li et al., 2021, Azar et al., 2013). The goal of this work is to bridge the gap between practical success of Q-learning and the relatively pessimistic theoretical results. The starting point of our work is the observation that in practice, Q-learning is used with two important modifications: (i) training with two networks, called online network and target network simultaneously (online target learning, or OTL) , and (ii) experience replay (ER) (Mnih et al., 2015). While they have been observed to play a significant role in the practical success of Q-learning, a thorough theoretical understanding of how these two modifications improve the convergence behavior of Q-learning has been missing in literature. By carefully combining the Q-learning with OTL and reverse experience replay (RER) (a form of experience replay), we present novel methods Q-Rex and Q-RexDaRe (Q-Rex+data reuse). We show that Q-Rex efficiently finds the optimal policy for linear MDPs and provide non-asymptotic bounds on sample complexity -- the first such result for a Q-learning method with linear MDPs. Furthermore, we demonstrate that Q-RexDaRe in fact achieves near optimal sample complexity in the tabular setting, improving upon the existing results for vanilla Q-learning.

Differentially Private Fractional Frequency Moments Estimation with Polylogarithmic Space

- Lun Wang, Iosif Pinelis, Dawn Song
- abstract@open-review(Poster): We prove that $\mathbb{F}_p$ sketch, a well-celebrated streaming algorithm for frequency moments estimation, is differentially private as is when $p\in(0, 1]$. $\mathbb{F}_p$ sketch uses only polylogarithmic space, exponentially better than existing DP baselines and only worse than the optimal non-private baseline by a logarithmic factor. The evaluation shows that $\mathbb{F}_p$ sketch can achieve reasonable accuracy with strong privacy guarantees. The code for evaluation is included in the supplementary material.

How Low Can We Go: Trading Memory for Error in Low-Precision Training

- Chengrun Yang, Ziyang Wu, Jerry Chee, Christopher De Sa, Madeleine Udell
- abstract@open-review(Poster): Low-precision arithmetic trains deep learning models using less energy, less memory and less time. However, we pay a price for the savings: lower precision may yield larger round-off error and hence larger prediction error. As applications proliferate, users must choose which precision to use to train a new model, and chip manufacturers must decide which precisions to manufacture. We view these precision choices as a hyperparameter tuning problem, and borrow ideas from meta-learning to learn the tradeoff between memory and error. In this paper, we introduce Pareto Estimation to Pick the Perfect Precision (PEPPP). We use matrix factorization to find non-dominated configurations (the Pareto frontier) with a limited number of network evaluations. For any given memory budget, the precision that minimizes error is a point on this frontier. Practitioners can use the frontier to trade memory for error and choose the best precision for their goals.

In a Nutshell, the Human Asked for This: Latent Goals for Following Temporal Specifications

- Borja G. LeÃ³n, Murray Shanahan, Francesco Belardinelli
- abstract@open-review(Poster): We address the problem of building agents whose goal is to learn to execute out-of distribution (OOD) multi-task instructions expressed in temporal logic (TL) by using deep reinforcement learning (DRL). Recent works provided evidence that the agent's neural architecture is a key feature when DRL agents are learning to solve OOD tasks in TL. Yet, the studies on this topic are still in their infancy. In this work, we propose a new deep learning configuration with inductive biases that lead agents to generate latent representations of their current goal, yielding a stronger generalization performance. We use these latent-goal networks within a neuro-symbolic framework that executes multi-task formally-defined

instructions and contrast the performance of the proposed neural networks against employing different state-of-the-art (SOTA) architectures when generalizing to unseen instructions in OOD environments.

Discrete Representations Strengthen Vision Transformer Robustness

- Chengzhi Mao, Lu Jiang, Mostafa Dehghani, Carl Vondrick, Rahul Sukthankar, Irfan Essa
- abstract@open-review(Poster): Vision Transformer (ViT) is emerging as the state-of-the-art architecture for image recognition. While recent studies suggest that ViTs are more robust than their convolutional counterparts, our experiments find that ViTs are overly reliant on local features (\eg, nuisances and texture) and fail to make adequate use of global context (\eg, shape and structure). As a result, ViTs fail to generalize to out-of-distribution, real-world data. To address this deficiency, we present a simple and effective architecture modification to ViT's input layer by adding discrete tokens produced by a vector-quantized encoder. Different from the standard continuous pixel tokens, discrete tokens are invariant under small perturbations and contain less information individually, which promote ViTs to learn global information that is invariant. Experimental results demonstrate that adding discrete representation on four architecture variants strengthens ViT robustness by up to 12\% across seven ImageNet robustness benchmarks while maintaining the performance on ImageNet.

On the Convergence of the Monte Carlo Exploring Starts Algorithm for Reinforcement Learning

- Che Wang, Shuhan Yuan, Kai Shao, Keith W. Ross
- abstract@open-review(Poster): A simple and natural algorithm for reinforcement learning (RL) is Monte Carlo Exploring Starts (MCES), where the Q-function is estimated by averaging the Monte Carlo returns, and the policy is improved by choosing actions that maximize the current estimate of the Q-function. Exploration is performed by "exploring starts", that is, each episode begins with a randomly chosen state and action, and then follows the current policy to the terminal state. In the classic book on RL by Sutton & Barto (2018), it is stated that establishing convergence for the MCES algorithm is one of the most important remaining open theoretical problems in RL. However, the convergence question for MCES turns out to be quite nuanced. Bertsekas & Tsitsiklis (1996) provide a counter-example showing that the MCES algorithm does not necessarily converge. Tsitsiklis (2002) further shows that if the original MCES algorithm is modified so that the Q-function estimates are updated at the same rate for all state-action pairs, and the discount factor is strictly less than one, then the MCES algorithm converges. In this paper we make headway with the original and more efficient MCES algorithm given in Sutton et al. (1998), establishing almost sure convergence for Optimal Policy Feed-Forward MDPs, which are MDPs whose states are not revisited within any episode when using an optimal policy. Such MDPs include a large class of environments such as all deterministic environments and all episodic environments with a timestep or any monotonically changing values as part of the state. Different from the previous proofs using stochastic approximations, we introduce a novel inductive approach, which is very simple and only makes use of the strong law of large numbers.

Concurrent Adversarial Learning for Large-Batch Training

- Yong Liu, Xiangning Chen, Minhao Cheng, Cho-Jui Hsieh, Yang You
- abstract@open-review(Poster): Large-batch training has become a commonly used technique when training neural networks with a large number of GPU/TPU processors. As batch size increases, stochastic optimizers tend to converge to sharp local minima, leading to degraded test performance. Current methods usually use extensive data augmentation to increase the batch size, but we found the performance gain with data augmentation decreases as batch size increases, and data augmentation will become insufficient after certain point. In this paper, we propose to use adversarial learning to increase the batch size in large-batch training. Despite being a natural choice for smoothing the decision surface and biasing towards a flat region, adversarial learning has not been successfully applied in large-batch training since it requires at least two sequential gradient computations at each step, which will at least double the running time compared with vanilla training even with a large number of processors. To overcome this issue, we propose a novel Concurrent Adversarial Learning (ConAdv) method that decouple the sequential gradient computations in adversarial learning by utilizing staled parameters. Experimental results demonstrate that ConAdv can successfully increase the batch size on both ResNet-50 and EfficientNet training on ImageNet while maintaining high accuracy. In particular, we show ConAdv along can achieve 75.3\% top-1 accuracy on ImageNet ResNet-50 training with 96K batch size, and the accuracy can be further improved to 76.2\% when combining ConAdv with data augmentation. This is the first work successfully scales ResNet-50 training batch size to 96K.

Multiset-Equivariant Set Prediction with Approximate Implicit Differentiation

- Yan Zhang, David W Zhang, Simon Lacoste-Julien, Gertjan J. Burghouts, Cees G. M. Snoek
- abstract@open-review(Poster): Most set prediction models in deep learning use set-equivariant operations, but they actually operate on multisets. We show that set-equivariant functions cannot represent certain functions on multisets, so we introduce the more appropriate notion of multiset-equivariance. We identify that the existing Deep Set Prediction Network (DSPN) can be multiset-equivariant without being hindered by set-equivariance and improve it with approximate implicit differentiation, allowing for better optimization while being faster and saving memory. In a range of toy experiments, we show that the perspective of multiset-equivariance is beneficial and that our changes to DSPN achieve better results in most cases. On CLEVR object property prediction, we substantially improve over the state-of-the-art Slot Attention from 8% to 77% in one of the strictest evaluation metrics because of the benefits made possible by implicit differentiation.

Learned Simulators for Turbulence

- Kim Stachenfeld, Drummond Buschman Fielding, Dmitrii Kochkov, Miles Cranmer, Tobias Pfaff, Jonathan Godwin, Can Cui, Shirley Ho, Peter Battaglia, Alvaro Sanchez-Gonzalez
- abstract@open-review(Poster): Turbulence simulation with classical numerical solvers requires high-resolution grids to accurately resolve dynamics. Here we train learned simulators at low spatial and temporal resolutions to capture turbulent dynamics generated at high resolution. We show that our proposed model can simulate turbulent dynamics more accurately than classical numerical solvers at the comparably low resolutions across various scientifically relevant metrics. Our model is trained end-to-end from data and is capable of learning a range of challenging chaotic and turbulent dynamics at low resolution, including trajectories generated by the state-of-the-art Athena++ engine. We show that our simpler, general-purpose architecture outperforms various more specialized, turbulence-specific architectures from the learned turbulence simulation literature. In general, we see that learned simulators yield unstable trajectories; however, we show that tuning training noise and temporal downsampling solves this problem. We also find that while generalization beyond the training distribution is a challenge for learned models, training noise, added loss constraints, and dataset augmentation can help. Broadly, we conclude that our learned simulator outperforms traditional solvers run on coarser grids, and emphasize that simple design choices can offer stability and robust generalization.

Modular Lifelong Reinforcement Learning via Neural Composition

- Jorge A Mendez, Harm van Seijen, ERIC EATON
- abstract@open-review(Poster): Humans commonly solve complex problems by decomposing them into easier subproblems and then combining the subproblem solutions. This type of compositional reasoning permits reuse of the subproblem solutions when tackling future tasks that share part of the underlying compositional structure. In a continual or lifelong reinforcement learning (RL) setting, this ability to decompose knowledge into reusable components would enable agents to quickly learn new RL tasks by leveraging accumulated compositional structures. We explore a particular form of composition based on neural modules and present a set of RL problems that intuitively admit compositional solutions. Empirically, we demonstrate that neural composition indeed captures the underlying structure of this space of problems. We further propose a compositional lifelong RL method that

leverages accumulated neural components to accelerate the learning of future tasks while retaining performance on previous tasks via off-line RL over replayed experiences.

Optimal ANN-SNN Conversion for High-accuracy and Ultra-low-latency Spiking Neural Networks

- Tong Bu, Wei Fang, Jianhao Ding, PENGLIN DAI, Zhaofei Yu, Tiejun Huang
- abstract@open-review(Poster): Spiking Neural Networks (SNNs) have gained great attraction due to their distinctive properties of low power consumption and fast inference on neuromorphic hardware. As the most effective method to get deep SNNs, ANN-SNN conversion has achieved comparable performance as ANNs on large-scale datasets. Despite this, it requires long time-steps to match the firing rates of SNNs to the activation of ANNs. As a result, the converted SNN suffers severe performance degradation problems with short time-steps, which hamper the practical application of SNNs. In this paper, we theoretically analyze ANN-SNN conversion error and derive the estimated activation function of SNNs. Then we propose the quantization clip-floor-shift activation function to replace the ReLU activation function in source ANNs, which can better approximate the activation function of SNNs. We prove that the expected conversion error between SNNs and ANNs is zero, enabling us to achieve high-accuracy and ultra-low-latency SNNs. We evaluate our method on CIFAR-10/100 and ImageNet datasets, and show that it outperforms the state-of-the-art ANN-SNN and directly trained SNNs in both accuracy and time-steps. To the best of our knowledge, this is the first time to explore high-performance ANN-SNN conversion with ultra-low latency (4 time-steps). Code is available at https://github.com/putshua/SNN_conversion_QCFS

AS-MLP: An Axial Shifted MLP Architecture for Vision

- Dongze Lian, Zehao Yu, Xing Sun, Shenghua Gao
- abstract@open-review(Poster): An Axial Shifted MLP architecture (AS-MLP) is proposed in this paper. Different from MLP-Mixer, where the global spatial feature is encoded for information flow through matrix transposition and one token-mixing MLP, we pay more attention to the local features interaction. By axially shifting channels of the feature map, AS-MLP is able to obtain the information flow from different axial directions, which captures the local dependencies. Such an operation enables us to utilize a pure MLP architecture to achieve the same local receptive field as CNN-like architecture. We can also design the receptive field size and dilation of blocks of AS-MLP, \emph{etc}, in the same spirit of convolutional neural networks. With the proposed AS-MLP architecture, our model obtains 83.3\% Top-1 accuracy with 88M parameters and 15.2 GFLOPs on the ImageNet-1K dataset. Such a simple yet effective architecture outperforms all MLP-based architectures and achieves competitive performance compared to the transformer-based architectures (\emph{e.g.}, Swin Transformer) even with slightly lower FLOPs. In addition, AS-MLP is also the first MLP-based architecture to be applied to the downstream tasks (\emph{e.g.}, object detection and semantic segmentation). The experimental results are also impressive. Our proposed AS-MLP obtains 51.5 mAP on the COCO validation set and 49.5 MS mIoU on the ADE20K dataset, which is competitive compared to the transformer-based architectures. Our AS-MLP establishes a strong baseline of MLP-based architecture. Code is available at \url{https://github.com/svip-lab/AS-MLP}.

Online Continual Learning on Class Incremental Blurry Task Configuration with Anytime Inference

- Hyunseo Koh, Dahyun Kim, Jung-Woo Ha, Jonghyun Choi
- abstract@open-review(Poster): Despite rapid advances in continual learning, a large body of research is devoted to improving performance in the existing setups. While a handful of work do propose new continual learning setups, they still lack practicality in certain aspects. For better practicality, we first propose a novel continual learning setup that is online, task-free, class-incremental, of blurry task boundaries and subject to inference queries at any moment. We additionally propose a new metric to better measure the performance of the continual learning methods subject to inference queries at any moment. To address the challenging setup and evaluation protocol, we propose an effective method that employs a new memory management scheme and novel learning techniques. Our empirical validation demonstrates that the proposed method outperforms prior arts by large margins. Code and data splits are available at https://github.com/naver-ai/i-Blurry.

Learning with Noisy Labels Revisited: A Study Using Real-World Human Annotations

- Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, Yang Liu
- abstract@open-review(Poster): Existing research on learning with noisy labels mainly focuses on synthetic label noise. The synthetic noise, though has clean structures which greatly enabled statistical analyses, often fails to model the real-world noise patterns. The recent literature has observed several efforts to offer real-world noisy datasets, e.g., Food-101N, WebVision, and Clothing1M. Yet the existing efforts suffer from two caveats: firstly, the lack of ground-truth verification makes it hard to theoretically study the property and treatment of real-world label noise. Secondly, these efforts are often of large scales, which may result in unfair comparisons of robust methods within reasonable and accessible computation power. To better understand real-world label noise, it is important to establish controllable, easy-to-use, and moderate-sized real-world noisy datasets with both ground-truth and noisy labels. This work presents two new benchmark datasets, which we name as CIFAR-10N, CIFAR-100N (jointly we call them CIFAR-N), equipping the training datasets of CIFAR-10 and CIFAR-100 with human-annotated real-world noisy labels we collected from Amazon Mechanical Turk. We quantitatively and qualitatively show that real-world noisy labels follow an instance-dependent pattern rather than the classically assumed and adopted ones (e.g., class-dependent label noise). We then initiate an effort to benchmarking a subset of the existing solutions using CIFAR-10N and CIFAR-100N. We further proceed to study the memorization of correct and wrong predictions, which further illustrates the difference between human noise and class-dependent synthetic noise. We show indeed the real-world noise patterns impose new and outstanding challenges as compared to synthetic label noise. These observations require us to rethink the treatment of noisy labels, and we hope the availability of these two datasets would facilitate the development and evaluation of future learning with noisy label solutions. The corresponding datasets and the leaderboard are available at http://noisylabels.com.

Optimization inspired Multi-Branch Equilibrium Models

- Mingjie Li, Yisen Wang, Xingyu Xie, Zhouchen Lin
- abstract@open-review(Poster): Works have shown the strong connections between some implicit models and optimization problems. However, explorations on such relationships are limited. Most works pay attention to some common mathematical properties, such as sparsity. In this work, we propose a new type of implicit model inspired by the designing of the systems' hidden objective functions, called the Multi-branch Optimization induced Equilibrium networks~(MOptEqs). The model architecture is designed based on modelling the hidden objective function for the multi-resolution recognition task. Furthermore, we also propose a new strategy inspired by our understandings of the hidden objective function. In this manner, the proposed model can better utilize the hierarchical patterns for recognition tasks and retain the abilities for interpreting the whole structure as trying to obtain the minima of the problem's goal. Comparing with the state-of-the-art models, our MOptEqs not only enjoys better explainability but are also superior to MDEQ with less parameter consumption and better performance on practical tasks. Furthermore, we also implement various experiments to demonstrate the effectiveness of our new methods and explore the applicability of the model's hidden objective function.

Learning to Annotate Part Segmentation with Gradient Matching

- Yu Yang, Xiaotian Cheng, Hakan Bilen, Xiangyang Ji
- abstract@open-review(Poster): The success of state-of-the-art deep neural networks heavily relies on the presence of large-scale labelled datasets, which are extremely expensive and time-consuming to annotate. This paper focuses on tackling semi-supervised part segmentation tasks by generating high-quality images with a pre-trained GAN and labelling the generated images with an automatic annotator. In particular, we formulate the annotator learning as a learning-to-learn problem. Given a pre-trained GAN, the annotator learns to label object parts in a set of randomly generated images such that a part

segmentation model trained on these synthetic images with their predicted labels obtains low segmentation error on a small validation set of manually labelled images. We further reduce this nested-loop optimization problem to a simple gradient matching problem and efficiently solve it with an iterative algorithm. We show that our method can learn annotators from a broad range of labelled images including real images, generated images, and even analytically rendered images. Our method is evaluated with semi-supervised part segmentation tasks and significantly outperforms other semi-supervised competitors when the amount of labelled examples is extremely limited.

Vector-quantized Image Modeling with Improved VQGAN

- Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, Yonghui Wu
- abstract@open-review(Poster): Pretraining language models with next-token prediction on massive text corpora has delivered phenomenal zero-shot, few-shot, transfer learning and multi-tasking capabilities on both generative and discriminative language tasks. Motivated by this success, we explore a Vector-quantized Image Modeling (VIM) approach that involves pretraining a Transformer to predict rasterized image tokens autoregressively. The discrete image tokens are encoded from a learned Vision-Transformer-based VQGAN (ViT-VQGAN). We first propose multiple improvements over vanilla VQGAN from architecture to codebook learning, yielding better efficiency and reconstruction fidelity. The improved ViT-VQGAN further improves vector-quantized image modeling tasks, including unconditional, class-conditioned image generation and unsupervised representation learning. When trained on ImageNet at 256x256 resolution, we achieve Inception Score (IS) of 175.1 and Fr'echet Inception Distance (FID) of 4.17, a dramatic improvement over the vanilla VQGAN, which obtains 70.6 and 17.04 for IS and FID, respectively. Based on ViT-VQGAN and unsupervised pretraining, we further evaluate the pretrained Transformer by averaging intermediate features, similar to Image GPT (iGPT). This ImageNet-pretrained VIM-L significantly beats iGPT-L on linear-probe accuracy from 60.3% to 73.2% for a similar model size. ViM-L also outperforms iGPT-XL which is trained with extra web image data and larger model size.

Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation

- Ofir Press, Noah Smith, Mike Lewis
- abstract@open-review(Poster): Since the introduction of the transformer model by Vaswani et al. (2017), a fundamental question has yet to be answered: how does a model achieve extrapolation at inference time for sequences that are longer than it saw during training? We first show that extrapolation can be enabled by simply changing the position representation method, though we find that current methods do not allow for efficient extrapolation. We therefore introduce a simpler and more efficient position method, Attention with Linear Biases (ALiBi). ALiBi does not add positional embeddings to word embeddings; instead, it biases query-key attention scores with a penalty that is proportional to their distance. We show that this method trains a 1.3 billion parameter model on input sequences of length 1024 that extrapolates to input sequences of length 2048, achieving the same perplexity as a sinusoidal position embedding model trained on inputs of length 2048 but training 11% faster and using 11% less memory. ALiBi's inductive bias towards recency also leads it to outperform multiple strong position methods on the WikiText-103 benchmark.

Learning Representation from Neural Fisher Kernel with Low-rank Approximation

- Ruixiang ZHANG, Shuangfei Zhai, Etai Littwin, Joshua M. Susskind
- abstract@open-review(Poster): In this paper, we study the representation of neural networks from the view of kernels. We first define the Neural Fisher Kernel (NFK), which is the Fisher Kernel applied to neural networks. We show that NFK can be computed for both supervised and unsupervised learning models, which can serve as a unified tool for representation extraction. Furthermore, we show that practical NFKs exhibit low-rank structures. We then propose an efficient algorithm that computes a low-rank approximation of NFK, which scales to large datasets and networks. We show that the low-rank approximation of NFKs derived from unsupervised generative models and supervised learning models gives rise to high-quality compact representations of data, achieving competitive results on a variety of machine learning tasks.

Learning Temporally Causal Latent Processes from General Temporal Data

- Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, Kun Zhang
- abstract@open-review(Poster): Our goal is to recover time-delayed latent causal variables and identify their relations from measured temporal data. Estimating causally-related latent variables from observations is particularly challenging as the latent variables are not uniquely recoverable in the most general case. In this work, we consider both a nonparametric, nonstationary setting and a parametric setting for the latent processes and propose two provable conditions under which temporally causal latent processes can be identified from their nonlinear mixtures. We propose LEAP, a theoretically-grounded framework that extends Variational AutoEncoders (VAEs) by enforcing our conditions through proper constraints in causal process prior. Experimental results on various datasets demonstrate that temporally causal latent processes are reliably identified from observed variables under different dependency structures and that our approach considerably outperforms baselines that do not properly leverage history or nonstationarity information. This demonstrates that using temporal information to learn latent processes from their invertible nonlinear mixtures in an unsupervised manner, for which we believe our work is one of the first, seems promising even without sparsity or minimality assumptions.

The Rich Get Richer: Disparate Impact of Semi-Supervised Learning

- Zhaowei Zhu, Tianyi Luo, Yang Liu
- abstract@open-review(Poster): Semi-supervised learning (SSL) has demonstrated its potential to improve the model accuracy for a variety of learning tasks when the high-quality supervised data is severely limited. Although it is often established that the average accuracy for the entire population of data is improved, it is unclear how SSL fares with different sub-populations. Understanding the above question has substantial fairness implications when different sub-populations are defined by the demographic groups that we aim to treat fairly. In this paper, we reveal the disparate impacts of deploying SSL: the sub-population who has a higher baseline accuracy without using SSL (the "rich" one) tends to benefit more from SSL; while the sub-population who suffers from a low baseline accuracy (the "poor" one) might even observe a performance drop after adding the SSL module. We theoretically and empirically establish the above observation for a broad family of SSL algorithms, which either explicitly or implicitly use an auxiliary "pseudo-label". Experiments on a set of image and text classification tasks confirm our claims. We introduce a new metric, Benefit Ratio, and promote the evaluation of the fairness of SSL (Equalized Benefit Ratio). We further discuss how the disparate impact can be mitigated. We hope our paper will alarm the potential pitfall of using SSL and encourage a multifaceted evaluation of future SSL algorithms.

Neural Relational Inference with Node-Specific Information

- Ershad Banijamali
- abstract@open-review(Poster): Inferring interactions among entities is an important problem in studying dynamical systems, which greatly impacts the performance of downstream tasks, such as prediction. In this paper, we tackle the relational inference problem in a setting where each entity can potentially have a set of individualized information that other entities cannot have access to. Specifically, we represent the system using a graph in which the individualized information become node-specific information (NSI). We build our model in the framework of Neural Relation Inference (NRI), where the interaction among entities are uncovered using variational inference. We adopt NRI model to incorporate the individualized information by introducing private nodes in the graph that represent NSI. Such representation enables us to uncover more accurate relations among the agents and therefore leads to better performance on the downstream tasks. Our experiment results over real-world datasets validate the merit of our proposed algorithm.

Bregman Gradient Policy Optimization

- Feihu Huang, Shangqian Gao, Heng Huang
- abstract@open-review(Poster): In the paper, we design a novel Bregman gradient policy optimization framework for reinforcement learning based on Bregman divergences and momentum techniques. Specifically, we propose a Bregman gradient policy optimization (BGPO) algorithm based on the basic momentum technique and mirror descent iteration. Meanwhile, we further propose an accelerated Bregman gradient policy optimization (VR-BGPO) algorithm based on the variance reduced technique. Moreover, we provide a convergence analysis framework for our Bregman gradient policy optimization under the nonconvex setting. We prove that our BGPO achieves a sample complexity of $O(\epsilon^{-4})$ for finding ϵ-stationary policy only requiring one trajectory at each iteration, and our VR-BGPO reaches the best known sample complexity of $O(\epsilon^{-3})$, which also only requires one trajectory at each iteration. In particular, by using different Bregman divergences, our BGPO framework unifies many existing policy optimization algorithms such as the existing (variance reduced) policy gradient algorithms such as natural policy gradient algorithm. Extensive experimental results on multiple reinforcement learning tasks demonstrate the efficiency of our new algorithms.

A generalization of the randomized singular value decomposition

- Nicolas Boulle, Alex Townsend
- abstract@open-review(Poster): The randomized singular value decomposition (SVD) is a popular and effective algorithm for computing a near-best rank k approximation of a matrix A using matrix-vector products with standard Gaussian vectors. Here, we generalize the theory of randomized SVD to multivariate Gaussian vectors, allowing one to incorporate prior knowledge of A into the algorithm. This enables us to explore the continuous analogue of the randomized SVD for Hilbert--Schmidt (HS) operators using operator-function products with functions drawn from a Gaussian process (GP). We then construct a new covariance kernel for GPs, based on weighted Jacobi polynomials, which allows us to rapidly sample the GP and control the smoothness of the randomly generated functions. Numerical examples on matrices and HS operators demonstrate the applicability of the algorithm.

Dropout Q-Functions for Doubly Efficient Reinforcement Learning

- Takuya Hiraoka, Takahisa Imagawa, Taisei Hashimoto, Takashi Onishi, Yoshimasa Tsuruoka
- abstract@open-review(Poster): Randomized ensembled double Q-learning (REDQ) (Chen et al., 2021b) has recently achieved state-of-the-art sample efficiency on continuous-action reinforcement learning benchmarks. This superior sample efficiency is made possible by using a large Q-function ensemble. However, REDQ is much less computationally efficient than non-ensemble counterparts such as Soft Actor-Critic (SAC) (Haarnoja et al., 2018a). To make REDQ more computationally efficient, we propose a method of improving computational efficiency called DroQ, which is a variant of REDQ that uses a small ensemble of dropout Q-functions. Our dropout Q-functions are simple Q-functions equipped with dropout connection and layer normalization. Despite its simplicity of implementation, our experimental results indicate that DroQ is doubly (sample and computationally) efficient. It achieved comparable sample efficiency with REDQ, much better computational efficiency than REDQ, and comparable computational efficiency with that of SAC.

QDrop: Randomly Dropping Quantization for Extremely Low-bit Post-Training Quantization

- Xiuying Wei, Ruihao Gong, Yuhang Li, Xianglong Liu, Fengwei Yu
- abstract@open-review(Poster): Recently, post-training quantization (PTQ) has driven much attention to produce efficient neural networks without long-time retraining. Despite the low cost, current PTQ works always fail under the extremely low-bit setting. In this study, we pioneeringly confirm that properly incorporating activation quantization into the PTQ reconstruction benefits the final accuracy. To deeply understand the inherent reason, a theoretical framework is established, which inspires us that the flatness of the optimized low-bit model on calibration and test data is crucial. Based on the conclusion, a simple yet effective approach dubbed as \textsc{QDrop} is proposed, which randomly drops the quantization of activations during reconstruction. Extensive experiments on various tasks including computer vision (image classification, object detection) and natural language processing (text classification and question answering) prove its superiority. With \textsc{QDrop}, the limit of PTQ is pushed to the 2-bit activation for the first time and the accuracy boost can be up to 51.49\%. Without bells and whistles, \textsc{QDrop} establishes a new state of the art for PTQ.

You Mostly Walk Alone: Analyzing Feature Attribution in Trajectory Prediction

- Osama Makansi, Julius Von KÃ¼gelgen, Francesco Locatello, Peter Vincent Gehler, Dominik Janzing, Thomas Brox, Bernhard SchÃ¶lkopf
- abstract@open-review(Poster): Predicting the future trajectory of a moving agent can be easy when the past trajectory continues smoothly but is challenging when complex interactions with other agents are involved. Recent deep learning approaches for trajectory prediction show promising performance and partially attribute this to successful reasoning about agent-agent interactions. However, it remains unclear which features such black-box models actually learn to use for making predictions. This paper proposes a procedure that quantifies the contributions of different cues to model performance based on a variant of Shapley values. Applying this procedure to state-of-the-art trajectory prediction methods on standard benchmark datasets shows that they are, in fact, unable to reason about interactions. Instead, the past trajectory of the target is the only feature used for predicting its future. For a task with richer social interaction patterns, on the other hand, the tested models do pick up such interactions to a certain extent, as quantified by our feature attribution method. We discuss the limits of the proposed method and its links to causality.

Rethinking Class-Prior Estimation for Positive-Unlabeled Learning

- Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Gang Niu, Masashi Sugiyama, Dacheng Tao
- abstract@open-review(Poster): Given only positive (P) and unlabeled (U) data, PU learning can train a binary classifier without any negative data. It has two building blocks: PU class-prior estimation (CPE) and PU classification; the latter has been well studied while the former has received less attention. Hitherto, the distributional-assumption-free CPE methods rely on a critical assumption that the support of the positive data distribution cannot be contained in the support of the negative data distribution. If this is violated, those CPE methods will systematically overestimate the class prior; it is even worse that we cannot verify the assumption based on the data. In this paper, we rethink CPE for PU learningâ€”can we remove the assumption to make CPE always valid? We show an affirmative answer by proposing Regrouping CPE (ReCPE) that builds an auxiliary probability distribution such that the support of the positive data distribution is never contained in the support of the negative data distribution. ReCPE can work with any CPE method by treating it as the base method. Theoretically, ReCPE does not affect its base if the assumption already holds for the original probability distribution; otherwise, it reduces the positive bias of its base. Empirically, ReCPE improves all state-of-the-art CPE methods on various datasets, implying that the assumption has indeed been violated here.

Learning Efficient Online 3D Bin Packing on Packing Configuration Trees

- Hang Zhao, Yang Yu, Kai Xu
- abstract@open-review(Poster): Online 3D Bin Packing Problem (3D-BPP) has widespread applications in industrial automation and has aroused enthusiastic research interest recently. Existing methods usually solve the problem with limited resolution of spatial discretization, and/or cannot deal with complex practical constraints well. We propose to enhance the practical applicability of online 3D-BPP via learning on a novel hierarchical representation â€“ packing configuration tree (PCT). PCT is a full-fledged description of the state and action space of bin packing which can support packing policy learning based on deep reinforcement learning (DRL). The size of the packing action space is proportional to the number of leaf nodes, making the DRL

model easy to train and well-performing even with continuous solution space. During training, PCT expands based on heuristic rules, however, the DRL model learns a much more effective and robust packing policy than heuristic methods. Through extensive evaluation, we demonstrate that our method outperforms all existing online BPP methods and is versatile in terms of incorporating various practical constraints.

Uncertainty Modeling for Out-of-Distribution Generalization

- Xiaotong Li, Yongxing Dai, Yixiao Ge, Jun Liu, Ying Shan, LINGYU DUAN
- abstract@open-review(Poster): Though remarkable progress has been achieved in various vision tasks, deep neural networks still suffer obvious performance degradation when tested in out-of-distribution scenarios. We argue that the feature statistics (mean and standard deviation), which carry the domain characteristics of the training data, can be properly manipulated to improve the generalization ability of deep learning models. Common methods often consider the feature statistics as deterministic values measured from the learned features and do not explicitly consider the uncertain statistics discrepancy caused by potential domain shifts during testing. In this paper, we improve the network generalization ability by modeling the uncertainty of domain shifts with synthesized feature statistics during training. Specifically, we hypothesize that the feature statistic, after considering the potential uncertainties, follows a multivariate Gaussian distribution. Hence, each feature statistic is no longer a deterministic value, but a probabilistic point with diverse distribution possibilities. With the uncertain feature statistics, the models can be trained to alleviate the domain perturbations and achieve better robustness against potential domain shifts. Our method can be readily integrated into networks without additional parameters. Extensive experiments demonstrate that our proposed method consistently improves the network generalization ability on multiple vision tasks, including image classification, semantic segmentation, and instance retrieval.

Online Adversarial Attacks

- Andjela Mladenovic, Joey Bose, Hugo berard, William L. Hamilton, Simon Lacoste-Julien, Pascal Vincent, Gauthier Gidel
- abstract@open-review(Poster): Adversarial attacks expose important vulnerabilities of deep learning models, yet little attention has been paid to settings where data arrives as a stream. In this paper, we formalize the online adversarial attack problem, emphasizing two key elements found in real-world use-cases: attackers must operate under partial knowledge of the target model, and the decisions made by the attacker are irrevocable since they operate on a transient data stream. We first rigorously analyze a deterministic variant of the online threat model by drawing parallels to the well-studied k-secretary problem in theoretical computer science and propose Virtual+, a simple yet practical online algorithm. Our main theoretical result shows Virtual+ yields provably the best competitive ratio over all single-threshold algorithms for $k<5$---extending the previous analysis of the k-secretary problem. We also introduce the \textit{stochastic k-secretary}---effectively reducing online blackbox transfer attacks to a k-secretary problem under noise---and prove theoretical bounds on the performance of Virtual+ adapted to this setting. Finally, we complement our theoretical results by conducting experiments on MNIST, CIFAR-10, and Imagenet classifiers, revealing the necessity of online algorithms in achieving near-optimal performance and also the rich interplay between attack strategies and online attack selection, enabling simple strategies like FGSM to outperform stronger adversaries.

Anytime Dense Prediction with Confidence Adaptivity

- Zhuang Liu, Zhiqiu Xu, Hung-Ju Wang, Trevor Darrell, Evan Shelhamer
- abstract@open-review(Poster): Anytime inference requires a model to make a progression of predictions which might be halted at any time. Prior research on anytime visual recognition has mostly focused on image classification.We propose the first unified and end-to-end approach for anytime dense prediction. A cascade of "exits" is attached to the model to make multiple predictions. We redesign the exits to account for the depth and spatial resolution of the features for each exit. To reduce total computation, and make full use of prior predictions, we develop a novel spatially adaptive approach to avoid further computation on regions where early predictions are already sufficiently confident. Our full method, named anytime dense prediction with confidence (ADP-C), achieves the same level of final accuracy, and meanwhile significantly reduces total computation. We evaluate our method on Cityscapes semantic segmentation and MPII human pose estimation: ADP-C enables anytime inference without sacrificing accuracy while also reducing the total FLOPs of its base models by 44.4% and 59.1%. We compare with anytime inference by deep equilibrium networks and feature-based stochastic sampling, showing that ADP-C dominates both across the accuracy-computation curve. Our code is available at https://github.com/liuzhuang13/anytime.

Declarative nets that are equilibrium models

- Russell Tsuchida, Suk Yee Yong, Mohammad Ali Armin, Lars Petersson, Cheng Soon Ong
- abstract@open-review(Poster): Implicit layers are computational modules that output the solution to some problem depending on the input and the layer parameters. Deep equilibrium models (DEQs) output a solution to a fixed point equation. Deep declarative networks (DDNs) solve an optimisation problem in their forward pass, an arguably more intuitive, interpretable problem than finding a fixed point. We show that solving a kernelised regularised maximum likelihood estimate as an inner problem in a DDN yields a large class of DEQ architectures. Our proof uses the exponential family in canonical form, and provides a closed-form expression for the DEQ parameters in terms of the kernel. The activation functions have interpretations in terms of the derivative of the log partition function. Building on existing literature, we interpret DEQs as fine-tuned, unrolled classical algorithms, giving an intuitive justification for why DEQ models are sensible. We use our theoretical result to devise an initialisation scheme for DEQs that allows them to solve kGLMs in their forward pass at initialisation. We empirically show that this initialisation scheme improves training stability and performance over random initialisation.

A Reduction-Based Framework for Conservative Bandits and Reinforcement Learning

- Yunchang Yang, Tianhao Wu, Han Zhong, Evrard Garcelon, Matteo Pirotta, Alessandro Lazaric, Liwei Wang, Simon Shaolei Du
- abstract@open-review(Poster): We study bandits and reinforcement learning (RL) subject to a conservative constraint where the agent is asked to perform at least as well as a given baseline policy. This setting is particular relevant in real-world domains including digital marketing, healthcare, production, finance, etc. In this paper, we present a reduction-based framework for conservative bandits and RL, in which our core technique is to calculate the necessary and sufficient budget obtained from running the baseline policy. For lower bounds, we improve the existing lower bound for conservative multi-armed bandits and obtain new lower bounds for conservative linear bandits, tabular RL and low-rank MDP, through a black-box reduction that turns a certain lower bound in the nonconservative setting into a new lower bound in the conservative setting. For upper bounds, in multi-armed bandits, linear bandits and tabular RL, our new upper bounds tighten or match existing ones with significantly simpler analyses. We also obtain a new upper bound for conservative low-rank MDP.

Wisdom of Committees: An Overlooked Approach To Faster and More Accurate Models

- Xiaofang Wang, Dan Kondratyuk, Eric Christiansen, Kris M. Kitani, Yair Movshovitz-Attias, Elad Eban
- abstract@open-review(Poster): Committee-based models (ensembles or cascades) construct models by combining existing pre-trained ones. While ensembles and cascades are well-known techniques that were proposed before deep learning, they are not considered a core building block of deep model architectures and are rarely compared to in recent literature on developing efficient models. In this work, we go back to basics and conduct a comprehensive analysis of the efficiency of committee-based models. We find that even the most simplistic method for building committees from existing, independently pre-trained models can match or exceed the accuracy of state-of-the-art models while being drastically more efficient. These simple committee-based models also outperform sophisticated neural architecture search methods (e.g., BigNAS). These findings hold true for several tasks, including image classification, video classification, and semantic segmentation, and various architecture families, such as ViT, EfficientNet, ResNet,

MobileNetV2, and X3D. Our results show that an EfficientNet cascade can achieve a 5.4x speedup over B7 and a ViT cascade can achieve a 2.3x speedup over ViT-L-384 while being equally accurate.

Unsupervised Discovery of Object Radiance Fields

- Hong-Xing Yu, Leonidas Guibas, Jiajun Wu
- abstract@open-review(Poster): We study the problem of inferring an object-centric scene representation from a single image, aiming to derive a representation that explains the image formation process, captures the scene's 3D nature, and is learned without supervision. Most existing methods on scene decomposition lack one or more of these characteristics, due to the fundamental challenge in integrating the complex 3D-to-2D image formation process into powerful inference schemes like deep networks. In this paper, we propose unsupervised discovery of Object Radiance Fields (uORF), integrating recent progresses in neural 3D scene representations and rendering with deep inference networks for unsupervised 3D scene decomposition. Trained on multi-view RGB images without annotations, uORF learns to decompose complex scenes with diverse, textured background from a single image. We show that uORF enables novel tasks, such as scene segmentation and editing in 3D, and it performs well on these tasks and on novel view synthesis on three datasets.

Gradient Step Denoiser for convergent Plug-and-Play

- Samuel Hurault, Arthur Leclaire, Nicolas Papadakis
- abstract@open-review(Poster): Plug-and-Play methods constitute a class of iterative algorithms for imaging problems where regularization is performed by an off-the-shelf denoiser. Although Plug-and-Play methods can lead to tremendous visual performance for various image problems, the few existing convergence guarantees are based on unrealistic (or suboptimal) hypotheses on the denoiser, or limited to strongly convex data terms. In this work, we propose a new type of Plug-and-Play methods, based on half-quadratic splitting, for which the denoiser is realized as a gradient descent step on a functional parameterized by a deep neural network. Exploiting convergence results for proximal gradient descent algorithms in the non-convex setting, we show that the proposed Plug-and-Play algorithm is a convergent iterative scheme that targets stationary points of an explicit global functional. Besides, experiments show that it is possible to learn such a deep denoiser while not compromising the performance in comparison to other state-of-the-art deep denoisers used in Plug-and-Play schemes. We apply our proximal gradient algorithm to various ill-posed inverse problems, e.g. deblurring, super-resolution and inpainting. For all these applications, numerical results empirically confirm the convergence results. Experiments also show that this new algorithm reaches state-of-the-art performance, both quantitatively and qualitatively.

Surrogate Gap Minimization Improves Sharpness-Aware Training

- Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha C Dvornek, sekhar tatikonda, James s Duncan, Ting Liu
- abstract@open-review(Poster): The recently proposed Sharpness-Aware Minimization (SAM) improves generalization by minimizing a perturbed loss defined as the maximum loss within a neighborhood in the parameter space. However, we show that both sharp and flat minima can have a low perturbed loss, implying that SAM does not always prefer flat minima. Instead, we define a surrogate gap, a measure equivalent to the dominant eigenvalue of Hessian at a local minimum when the radius of neighborhood (to derive the perturbed loss) is small. The surrogate gap is easy to compute and feasible for direct minimization during training. Based on the above observations, we propose Surrogate Gap Guided Sharpness-Aware Minimization (GSAM), a novel improvement over SAM with negligible computation overhead. Conceptually, GSAM consists of two steps: 1) a gradient descent like SAM to minimize the perturbed loss, and 2) an ascent step in the orthogonal direction (after gradient decomposition) to minimize the surrogate gap and yet not affect the perturbed loss. GSAM seeks a region with both small loss (by step 1) and low sharpness (by step 2), giving rise to a model with high generalization capabilities. Theoretically, we show the convergence of GSAM and provably better generalization than SAM.Empirically, GSAM consistently improves generalization (e.g., +3.2% over SAM and +5.4% over AdamW on ImageNet top-1 accuracy for ViT-B/32). Code is released at https://sites.google.com/view/gsam-iclr22/home

R4D: Utilizing Reference Objects for Long-Range Distance Estimation

- Yingwei Li, Tiffany Chen, Maya Kabkab, Ruichi Yu, Longlong Jing, Yurong You, Hang Zhao
- abstract@open-review(Poster): Estimating the distance of objects is a safety-critical task for autonomous driving. Focusing on short-range objects, existing methods and datasets neglect the equally important long-range objects. In this paper, we introduce a challenging and under-explored task, which we refer to as Long-Range Distance Estimation, as well as two datasets to validate new methods developed for this task. We then proposeR4D, the first framework to accurately estimate the distance of long-range objects by using references with known distances in the scene. Drawing inspiration from human perception, R4D builds a graph by connecting a target object to all references. An edge in the graph encodes the relative distance information between a pair of target and reference objects. An attention module is then used to weigh the importance of reference objects and combine them into one target object distance prediction. Experiments on the two proposed datasets demonstrate the effectiveness and robustness of R4D by showing significant improvements compared to existing baselines. Weâ€™re looking to make the proposed dataset, Waymo OpenDataset - Long-Range Labels, available publicly at waymo.com/open/download.

Understanding Dimensional Collapse in Contrastive Self-supervised Learning

- Li Jing, Pascal Vincent, Yann LeCun, Yuandong Tian
- abstract@open-review(Poster): Self-supervised visual representation learning aims to learn useful representations without relying on human annotations. Joint embedding approach bases on maximizing the agreement between embedding vectors from different views of the same image. Various methods have been proposed to solve the collapsing problem where all embedding vectors collapse to a trivial constant solution. Among these methods, contrastive learning prevents collapse via negative sample pairs. It has been shown that non-contrastive methods suffer from a lesser collapse problem of a different nature: dimensional collapse, whereby the embedding vectors end up spanning a lower-dimensional subspace instead of the entire available embedding space. Here, we show that dimensional collapse also happens in contrastive learning. In this paper, we shed light on the dynamics at play in contrastive learning that leads to dimensional collapse. Inspired by our theory, we propose a novel contrastive learning method, called DirectCLR, which directly optimizes the representation space without relying on a trainable projector. Experiments show that DirectCLR outperforms SimCLR with a trainable linear projector on ImageNet.

FedPara: Low-rank Hadamard Product for Communication-Efficient Federated Learning

- Nam Hyeon-Woo, Moon Ye-Bin, Tae-Hyun Oh
- abstract@open-review(Poster): In this work, we propose a communication-efficient parameterization, $\texttt{FedPara}$, for federated learning (FL) to overcome the burdens on frequent model uploads and downloads. Our method re-parameterizes weight parameters of layers using low-rank weights followed by the Hadamard product. Compared to the conventional low-rank parameterization, our $\texttt{FedPara}$ method is not restricted to low-rank constraints, and thereby it has a far larger capacity. This property enables to achieve comparable performance while requiring 3 to 10 times lower communication costs than the model with the original layers, which is not achievable by the traditional low-rank methods. The efficiency of our method can be further improved by combining with other efficient FL optimizers. In addition, we extend our method to a personalized FL application, $\texttt{pFedPara}$, which separates parameters into global and local ones. We show that $\texttt{pFedPara}$ outperforms competing personalized FL methods with more than three times fewer parameters.

RegionViT: Regional-to-Local Attention for Vision Transformers

- Chun-Fu Chen, Rameswar Panda, Quanfu Fan
- abstract@open-review(Poster): Vision transformer (ViT) has recently shown its strong capability in achieving comparable results to convolutional neural networks (CNNs) on image classification. However, vanilla ViT simply inherits the same architecture from the natural language processing directly, which is often not optimized for vision applications. Motivated by this, in this paper, we propose a new architecture that adopts the pyramid structure and employ novel regional-to-local attention rather than global self-attention in vision transformers. More specifically, our model first generates regional tokens and local tokens from an image with different patch sizes, where each regional token is associated with a set of local tokens based on the spatial location. The regional-to-local attention includes two steps: first, the regional self-attention extracts global information among all regional tokens and then the local self-attention exchanges the information among one regional token and the associated local tokens via self-attention. Therefore, even though local self-attention confines the scope in a local region but it can still receive global information. Extensive experiments on four vision tasks, including image classification, object and keypoint detection, semantics segmentation and action recognition, show that our approach outperforms or is on par with state-of-the-art ViT variants including many concurrent works. Our source codes and models are available at \url{https://github.com/IBM/RegionViT}.

Quadtree Attention for Vision Transformers

- Shitao Tang, Jiahui Zhang, Siyu Zhu, Ping Tan
- abstract@open-review(Poster): Transformers have been successful in many vision tasks, thanks to their capability of capturing long-range dependency. However, their quadratic computational complexity poses a major obstacle for applying them to vision tasks requiring dense predictions, such as object detection, feature matching, stereo, etc. We introduce QuadTree Attention, which reduces the computational complexity from quadratic to linear. Our quadtree transformer builds token pyramids and computes attention in a coarse-to-fine manner. At each level, the top K patches with the highest attention scores are selected, such that at the next level, attention is only evaluated within the relevant regions corresponding to these top K patches. We demonstrate that quadtree attention achieves state-of-the-art performance in various vision tasks, e.g. with 4.0% improvement in feature matching on ScanNet, about 50% flops reduction in stereo matching, 0.4-1.5% improvement in top-1 accuracy on ImageNet classification, 1.2-1.8% improvement on COCO object detection, and 0.7-2.4% improvement on semantic segmentation over previous state-of-the-art transformers. The codes are available at https://github.com/Tangshitao/QuadtreeAttention.

Visual Correspondence Hallucination

- Hugo Germain, Vincent Lepetit, Guillaume Bourmaud
- abstract@open-review(Poster): Given a pair of partially overlapping source and target images and a keypoint in the source image, the keypoint's correspondent in the target image can be either visible, occluded or outside the field of view. Local feature matching methods are only able to identify the correspondent's location when it is visible, while humans can also hallucinate its location when it is occluded or outside the field of view through geometric reasoning. In this paper, we bridge this gap by training a network to output a peaked probability distribution over the correspondent's location, regardless of this correspondent being visible, occluded, or outside the field of view. We experimentally demonstrate that this network is indeed able to hallucinate correspondences on pairs of images captured in scenes that were not seen at training-time. We also apply this network to an absolute camera pose estimation problem and find it is significantly more robust than state-of-the-art local feature matching-based competitors.

Whatâ€™s Wrong with Deep Learning in Tree Search for Combinatorial Optimization

- Maximilian BÃ¶ther, Otto KiÃŸig, Martin Taraz, Sarel Cohen, Karen Seidel, Tobias Friedrich
- abstract@open-review(Poster): Combinatorial optimization lies at the core of many real-world problems. Especially since the rise of graph neural networks (GNNs), the deep learning community has been developing solvers that derive solutions to NP-hard problems by learning the problem-specific solution structure. However, reproducing the results of these publications proves to be difficult. We make three contributions. First, we present an open-source benchmark suite for the NP-hard Maximum Independent Set problem, in both its weighted and unweighted variants. The suite offers a unified interface to various state-of-the-art traditional and machine learning-based solvers. Second, using our benchmark suite, we conduct an in-depth analysis of the popular guided tree search algorithm by Li et al. [NeurIPS 2018], testing various configurations on small and large synthetic and real-world graphs. By re-implementing their algorithm with a focus on code quality and extensibility, we show that the graph convolution network used in the tree search does not learn a meaningful representation of the solution structure, and can in fact be replaced by random values. Instead, the tree search relies on algorithmic techniques like graph kernelization to find good solutions. Thus, the results from the original publication are not reproducible. Third, we extend the analysis to compare the tree search implementations to other solvers, showing that the classical algorithmic solvers often are faster, while providing solutions of similar quality. Additionally, we analyze a recent solver based on reinforcement learning and observe that for this solver, the GNN is responsible for the competitive solution quality.

Deep Attentive Variational Inference

- Ifigeneia Apostolopoulou, Ian Char, Elan Rosenfeld, Artur Dubrawski
- abstract@open-review(Poster): Stochastic Variational Inference is a powerful framework for learning large-scale probabilistic latent variable models. However, typical assumptions on the factorization or independence of the latent variables can substantially restrict its capacity for inference and generative modeling. A major line of active research aims at building more expressive variational models by designing deep hierarchies of interdependent latent variables. Although these models exhibit superior performance and enable richer latent representations, we show that they incur diminishing returns: adding more stochastic layers to an already very deep model yields small predictive improvement while substantially increasing the inference and training time. Moreover, the architecture for this class of models favors local interactions among the latent variables between neighboring layers when designing the conditioning factors of the involved distributions. This is the first work that proposes attention mechanisms to build more expressive variational distributions in deep probabilistic models by explicitly modeling both local and global interactions in the latent space. Specifically, we propose deep attentive variational autoencoder and test it on a variety of established datasets. We show it achieves state-of-the-art log-likelihoods while using fewer latent layers and requiring less training time than existing models. The proposed non-local inference reduces computational footprint by alleviating the need for deep hierarchies. Project code: https://github.com/ifiaposto/Deep_Attentive_VI

ARTEMIS: Attention-based Retrieval with Text-Explicit Matching and Implicit Similarity

- Ginger Delmas, Rafael S. Rezende, Gabriela Csurka, Diane Larlus
- abstract@open-review(Poster): An intuitive way to search for images is to use queries composed of an example image and a complementary text. While the first provides rich and implicit context for the search, the latter explicitly calls for new traits, or specifies how some elements of the example image should be changed to retrieve the desired target image. Current approaches typically combine the features of each of the two elements of the query into a single representation, which can then be compared to the ones of the potential target images. Our work aims at shedding new light on the task by looking at it through the prism of two familiar and related frameworks: text-to-image and image-to-image retrieval. Taking inspiration from them, we exploit the specific relation of each query element with the targeted image and derive light-weight attention mechanisms which enable to mediate between the two complementary modalities. We validate our approach on several retrieval benchmarks, querying with images and their associated free-form text modifiers. Our method obtains state-of-the-art results without resorting to side information, multi-level features, heavy pre-training nor large architectures as in previous works.

Trivial or Impossible --- dichotomous data difficulty masks model differences (on ImageNet and beyond)

- Kristof Meding, Luca M. Schulze Buschoff, Robert Geirhos, Felix A. Wichmann
- abstract@open-review(Poster): "The power of a generalization system follows directly from its biases" (Mitchell 1980). Today, CNNs are incredibly powerful generalisation systems---but to what degree have we understood how their inductive bias influences model decisions? We here attempt to disentangle the various aspects that determine how a model decides. In particular, we ask: what makes one model decide differently from another? In a meticulously controlled setting, we find that (1.) irrespective of the network architecture or objective (e.g. self-supervised, semi-supervised, vision transformers, recurrent models) all models end up with a similar decision boundary. (2.) To understand these findings, we analysed model decisions on the ImageNet validation set from epoch to epoch and image by image. We find that the ImageNet validation set, among others, suffers from dichotomous data difficulty (DDD): For the range of investigated models and their accuracies, it is dominated by 46.0% "trivial" and 11.5% "impossible" images (beyond label errors). Only 42.5% of the images could possibly be responsible for the differences between two models' decision boundaries. (3.) Only removing the "impossible" and "trivial" images allows us to see pronounced differences between models. (4.) Humans are highly accurate at predicting which images are "trivial" and "impossible" for CNNs (81.4%). This implies that in future comparisons of brains, machines and behaviour, much may be gained from investigating the decisive role of images and the distribution of their difficulties.

Group equivariant neural posterior estimation

- Maximilian Dax, Stephen R Green, Jonathan Gair, Michael Deistler, Bernhard SchÃ¶lkopf, Jakob H. Macke
- abstract@open-review(Poster): Simulation-based inference with conditional neural density estimators is a powerful approach to solving inverse problems in science. However, these methods typically treat the underlying forward model as a black box, with no way to exploit geometric properties such as equivariances. Equivariances are common in scientific models, however integrating them directly into expressive inference networks (such as normalizing flows) is not straightforward. We here describe an alternative method to incorporate equivariances under joint transformations of parameters and data. Our method---called group equivariant neural posterior estimation (GNPE)---is based on self-consistently standardizing the "pose" of the data while estimating the posterior over parameters. It is architecture-independent, and applies both to exact and approximate equivariances. As a real-world application, we use GNPE for amortized inference of astrophysical binary black hole systems from gravitational-wave observations. We show that GNPE achieves state-of-the-art accuracy while reducing inference times by three orders of magnitude.

Fast Differentiable Matrix Square Root

- Yue Song, Nicu Sebe, Wei Wang
- abstract@open-review(Poster): Computing the matrix square root or its inverse in a differentiable manner is important in a variety of computer vision tasks. Previous methods either adopt the Singular Value Decomposition (SVD) to explicitly factorize the matrix or use the Newton-Schulz iteration (NS iteration) to derive the approximate solution. However, both methods are not computationally efficient enough in either the forward pass or in the backward pass. In this paper, we propose two more efficient variants to compute the differentiable matrix square root. For the forward propagation, one method is to use Matrix Taylor Polynomial (MTP), and the other method is to use Matrix Pad\'e Approximants (MPA). The backward gradient is computed by iteratively solving the continuous-time Lyapunov equation using the matrix sign function. Both methods yield considerable speed-up compared with the SVD or the Newton-Schulz iteration. Experimental results on the de-correlated batch normalization and second-order vision transformer demonstrate that our methods can also achieve competitive and even slightly better performances. The code is available at \href{https://github.com/KingJamesSong/FastDifferentiableMatSqrt}{https://github.com/KingJamesSong/FastDifferentiableMatSqrt}.

SQuant: On-the-Fly Data-Free Quantization via Diagonal Hessian Approximation

- Cong Guo, Yuxian Qiu, Jingwen Leng, Xiaotian Gao, Chen Zhang, Yunxin Liu, Fan Yang, Yuhao Zhu, Minyi Guo
- abstract@open-review(Poster): Quantization of deep neural networks (DNN) has been proven effective for compressing and accelerating DNN models. Data-free quantization (DFQ) is a promising approach without the original datasets under privacy-sensitive and confidential scenarios. However, current DFQ solutions degrade accuracy, need synthetic data to calibrate networks, and are time-consuming and costly. This paper proposes an on-the-fly DFQ framework with sub-second quantization time, called SQuant, which can quantize networks on inference-only devices with low computation and memory requirements. With the theoretical analysis of the second-order information of DNN task loss, we decompose and approximate the Hessian-based optimization objective into three diagonal sub-items, which have different areas corresponding to three dimensions of weight tensor: element-wise, kernel-wise, and output channel-wise. Then, we progressively compose sub-items and propose a novel data-free optimization objective in the discrete domain, minimizing Constrained Absolute Sum of Error (or CASE in short), which surprisingly does not need any dataset and is even not aware of network architecture. We also design an efficient algorithm without back-propagation to further reduce the computation complexity of the objective solver. Finally, without fine-tuning and synthetic datasets, SQuant accelerates the data-free quantization process to a sub-second level with >30% accuracy improvement over the existing data-free post-training quantization works, with the evaluated models under 4-bit quantization. We have open-sourced the SQuant framework at https://github.com/clevercool/SQuant.

Neural Variational Dropout Processes

- Insu Jeon, Youngjin Park, Gunhee Kim
- abstract@open-review(Poster): Learning to infer the conditional posterior model is a key step for robust meta-learning. This paper presents a new Bayesian meta-learning approach called Neural Variational Dropout Processes (NVDPs). NVDPs model the conditional posterior distribution based on a task-specific dropout; a low-rank product of Bernoulli experts meta-model is utilized for a memory-efficient mapping of dropout rates from a few observed contexts. It allows for a quick reconfiguration of a globally learned and shared neural network for new tasks in multi-task few-shot learning. In addition, NVDPs utilize a novel prior conditioned on the whole task data to optimize the conditional dropout posterior in the amortized variational inference. Surprisingly, this enables the robust approximation of task-specific dropout rates that can deal with a wide range of functional ambiguities and uncertainties. We compared the proposed method with other meta-learning approaches in the few-shot learning tasks such as 1D stochastic regression, image inpainting, and classification. The results show the excellent performance of NVDPs.

Towards Better Understanding and Better Generalization of Low-shot Classification in Histology Images with Contrastive Learning

- Jiawei Yang, Hanbo Chen, Jiangpeng Yan, Xiaoyu Chen, Jianhua Yao
- abstract@open-review(Poster): Few-shot learning is an established topic in natural images for years, but few work is attended to histology images, which is of high clinical value since well-labeled datasets and rare abnormal samples are expensive to collect. Here, we facilitate the study of few-shot learning in histology images by setting up three cross-domain tasks that simulate real clinics problems. To enable label-efficient learning and better generalizability, we propose to incorporate contrastive learning (CL) with latent augmentation (LA) to build a few-shot system. CL learns useful representations without manual labels, while LA transfers semantic variations of the base dataset in an unsupervised way. These two components fully exploit unlabeled training data and can scale gracefully to other label-hungry problems. In experiments, we find i) models learned by CL generalize better than supervised learning for histology images in unseen classes, and ii) LA brings consistent gains over baselines. Prior studies of self-supervised learning mainly focus on ImageNet-like images, which only present a dominant object in their centers. Recent attention has been paid to images with multi-objects and multi-textures. Histology images are a natural choice for such a study. We show the superiority of CL over supervised learning in terms of generalization for

such data and provide our empirical understanding for this observation. The findings in this work could contribute to understanding how the model generalizes in the context of both representation learning and histological image analysis. Code is available.

Distilling GANs with Style-Mixed Triplets for X2I Translation with Limited Data

- Yaxing Wang, Joost van de weijer, Lu Yu, SHANGLING JUI
- abstract@open-review(Poster): Conditional image synthesis is an integral part of many X2I translation systems, including image-to-image, text-to-image and audio-to-image translation systems. Training these large systems generally requires huge amounts of training data. Therefore, we investigate knowledge distillation to transfer knowledge from a high-quality unconditioned generative model (e.g., StyleGAN) to a conditioned synthetic image generation modules in a variety of systems. To initialize the conditional and reference branch (from a unconditional GAN) we exploit the style mixing characteristics of high-quality GANs to generate an infinite supply of style-mixed triplets to perform the knowledge distillation. Extensive experimental results in a number of image generation tasks (i.e., image-to-image, semantic segmentation-to-image, text-to-image and audio-to-image) demonstrate qualitatively and quantitatively that our method successfully transfers knowledge to the synthetic image generation modules, resulting in more realistic images than previous methods as confirmed by a significant drop in the FID.

Handling Distribution Shifts on Graphs: An Invariance Perspective

- Qitian Wu, Hengrui Zhang, Junchi Yan, David Wipf
- abstract@open-review(Poster): There is increasing evidence suggesting neural networks' sensitivity to distribution shifts, so that research on out-of-distribution (OOD) generalization comes into the spotlight. Nonetheless, current endeavors mostly focus on Euclidean data, and its formulation for graph-structured data is not clear and remains under-explored, given two-fold fundamental challenges: 1) the inter-connection among nodes in one graph, which induces non-IID generation of data points even under the same environment, and 2) the structural information in the input graph, which is also informative for prediction. In this paper, we formulate the OOD problem on graphs and develop a new invariant learning approach, Explore-to-Extrapolate Risk Minimization (EERM), that facilitates graph neural networks to leverage invariance principles for prediction. EERM resorts to multiple context explorers (specified as graph structure editers in our case) that are adversarially trained to maximize the variance of risks from multiple virtual environments. Such a design enables the model to extrapolate from a single observed environment which is the common case for node-level prediction. We prove the validity of our method by theoretically showing its guarantee of a valid OOD solution and further demonstrate its power on various real-world datasets for handling distribution shifts from artificial spurious features, cross-domain transfers and dynamic graph evolution.

Automatic Loss Function Search for Predict-Then-Optimize Problems with Strong Ranking Property

- Boshi Wang, Jialin Yi, Hang Dong, Bo Qiao, Chuan Luo, Qingwei Lin
- abstract@open-review(Poster): Combinatorial optimization problems with parameters to be predicted from side information are commonly seen in a variety of problems during the paradigm shift from reactive decision making to proactive decision making. Due to the misalignment between the continuous prediction results and the discrete decisions in optimization problems, it is hard to achieve a satisfactory prediction result with the ordinary l_2 loss in the prediction phase. To properly connect the prediction loss with the optimization goal, in this paper we propose a total group preorder (TGP) loss and its differential version called approximated total group preorder (ATGP) loss for predict-then-optimize (PTO) problems with strong ranking property. These new losses are provably more robust than the usual l_2 loss in a linear regression setting and have great potential to extend to other settings. We also propose an automatic searching algorithm that adapts the ATGP loss to PTO problems with different combinatorial structures. Extensive experiments on the ranking problem, the knapsack problem, and the shortest path problem have demonstrated that our proposed method can achieve a significant performance compared to the other methods designed for PTO problems.

Generalized Demographic Parity for Group Fairness

- Zhimeng Jiang, Xiaotian Han, Chao Fan, Fan Yang, Ali Mostafavi, Xia Hu
- abstract@open-review(Poster): This work aims to generalize demographic parity to continuous sensitive attributes while preserving tractable computation. Current fairness metrics for continuous sensitive attributes largely rely on intractable statistical independence between variables, such as Hirschfeld-Gebelein-Renyi (HGR) and mutual information. Statistical fairness metrics estimation relying on either tractable bounds or neural network approximation, however, are not sufficiently trustful to rank algorithms prediction bias due to lack of estimation accuracy guarantee. To make fairness metrics trustable, we propose \textit{\underline{G}eneralized \underline{D}emographic \underline{P}arity} (GDP), a group fairness metric for continuous and discrete attributes. We show the understanding of GDP from the probability perspective and theoretically reveal the connection between GDP regularizer and adversarial debiasing. To estimate GDP, we adopt hard and soft group strategies via the one-hot or the soft group indicator, representing the membership of each sample in different groups of the sensitive attribute. We provably and numerically show that the soft group strategy achieves a faster estimation error convergence rate. Experiments show the better bias mitigation performance of GDP regularizer, compared with adversarial debiasing, for regression and classification tasks in tabular and graph benchmarks.

Closed-form Sample Probing for Learning Generative Models in Zero-shot Learning

- Samet Cetin, Orhun BuÄŸra Baran, Ramazan Gokberk Cinbis
- abstract@open-review(Poster): Generative model based approaches have led to significant advances in zero-shot learning (ZSL) over the past few years. These approaches typically aim to learn a conditional generator that synthesizes training samples of classes conditioned on class definitions. The final zero-shot learning model is then obtained by training a supervised classification model over the real and/or synthesized training samples of seen and unseen classes, combined. Therefore, naturally, the generative model needs to produce not only relevant samples, but also those that are sufficiently rich for classifier training purposes, which is handled by various heuristics in existing works. In this paper, we introduce a principled approach for training generative models {\em directly} for training data generation purposes. Our main observation is that the use of closed-form models opens doors to end-to-end training thanks to the differentiability of the solvers. In our approach, at each generative model update step, we fit a task-specific closed-form ZSL model from generated samples, and measure its loss on novel samples all within the compute graph, a procedure that we refer to as {\em sample probing}. In this manner, the generator receives feedback directly based on the value of its samples for model training purposes. Our experimental results show that the proposed sample probing approach improves the ZSL results even when integrated into state-of-the-art generative models.

DKM: Differentiable k-Means Clustering Layer for Neural Network Compression

- Minsik Cho, Keivan Alizadeh-Vahid, Saurabh Adya, Mohammad Rastegari
- abstract@open-review(Poster): Deep neural network (DNN) model compression for efficient on-device inference is becoming increasingly important to reduce memory requirements and keep user data on-device. To this end, we propose a novel differentiable k-means clustering layer (DKM) and its application to train-time weight clustering-based DNN model compression. DKM casts k-means clustering as an attention problem and enables joint optimization of the DNN parameters and clustering centroids. Unlike prior works that rely on additional regularizers and parameters, DKM-based compression keeps the original loss function and model architecture fixed. We evaluated DKM-based compression on various DNN models for computer vision and natural language processing (NLP) tasks. Our results demonstrate that DKM delivers superior compression and accuracy trade-off on ImageNet1k and GLUE benchmarks. For example, DKM-based compression can offer 74.5% top-1 ImageNet1k accuracy on ResNet50 DNN model with 3.3MB model size (29.4x model compression factor). For MobileNet-v1, which is a challenging DNN to compress, DKM delivers 63.9% top-1 ImageNet1k accuracy with 0.72 MB model size (22.4x model compression factor). This result is 6.8% higher top-1accuracy and 33% relatively smaller

model size than the current state-of-the-art DNN compression algorithms. Additionally, DKM enables compression of DistilBERT model by 11.8x with minimal (1.1%) accuracy loss on GLUE NLP benchmarks.

Fixed Neural Network Steganography: Train the images, not the network

- Varsha Kishore, Xiangyu Chen, Yan Wang, Boyi Li, Kilian Q Weinberger
- abstract@open-review(Poster): Recent attempts at image steganography make use of advances in deep learning to train an encoder-decoder network pair to hide and retrieve secret messages in images. These methods are able to hide large amounts of data, but they also incur high decoding error rates (around 20%). In this paper, we propose a novel algorithm for steganography that takes advantage of the fact that neural networks are sensitive to tiny perturbations. Our method, Fixed Neural Network Steganography (FNNS), yields significantly lower error rates when compared to prior state-of-the-art methods and achieves 0% error reliably for hiding up to 3 bits per pixel (bpp) of secret information in images. FNNS also successfully evades existing statistical steganalysis systems and can be modified to evade neural steganalysis systems as well. Recovering every bit correctly, up to 3 bpp, enables novel applications that requires encryption. We introduce one specific use case for facilitating anonymized and safe image sharing. Our code is available at https://github.com/varshakishore/FNNS.

Steerable Partial Differential Operators for Equivariant Neural Networks

- Erik Jenner, Maurice Weiler
- abstract@open-review(Poster): Recent work in equivariant deep learning bears strong similarities to physics. Fields over a base space are fundamental entities in both subjects, as are equivariant maps between these fields. In deep learning, however, these maps are usually defined by convolutions with a kernel, whereas they are partial differential operators (PDOs) in physics. Developing the theory of equivariant PDOs in the context of deep learning could bring these subjects even closer together and lead to a stronger flow of ideas. In this work, we derive a G-steerability constraint that completely characterizes when a PDO between feature vector fields is equivariant, for arbitrary symmetry groups G. We then fully solve this constraint for several important groups. We use our solutions as equivariant drop-in replacements for convolutional layers and benchmark them in that role. Finally, we develop a framework for equivariant maps based on Schwartz distributions that unifies classical convolutions and differential operators and gives insight about the relation between the two.

Divergence-aware Federated Self-Supervised Learning

- Weiming Zhuang, Yonggang Wen, Shuai Zhang
- abstract@open-review(Poster): Self-supervised learning (SSL) is capable of learning remarkable representations from centrally available data. Recent works further implement federated learning with SSL to learn from rapidly growing decentralized unlabeled images (e.g., from cameras and phones), often resulted from privacy constraints. Extensive attention has been paid to SSL approaches based on Siamese networks. However, such an effort has not yet revealed deep insights into various fundamental building blocks for the federated self-supervised learning (FedSSL) architecture. We aim to fill in this gap via in-depth empirical study and propose a new method to tackle the non-independently and identically distributed (non-IID) data problem of decentralized data. Firstly, we introduce a generalized FedSSL framework that embraces existing SSL methods based on Siamese networks and presents flexibility catering to future methods. In this framework, a server coordinates multiple clients to conduct SSL training and periodically updates local models of clients with the aggregated global model. Using the framework, our study uncovers unique insights of FedSSL: 1) stop-gradient operation, previously reported to be essential, is not always necessary in FedSSL; 2) retaining local knowledge of clients in FedSSL is particularly beneficial for non-IID data. Inspired by the insights, we then propose a new approach for model update, Federated Divergence-aware Exponential Moving Average update (FedEMA). FedEMA updates local models of clients adaptively using EMA of the global model, where the decay rate is dynamically measured by model divergence. Extensive experiments demonstrate that FedEMA outperforms existing methods by 3-4% on linear evaluation. We hope that this work will provide useful insights for future research.

Neural Spectral Marked Point Processes

- Shixiang Zhu, Haoyun Wang, Zheng Dong, Xiuyuan Cheng, Yao Xie
- abstract@open-review(Poster): Self- and mutually-exciting point processes are popular models in machine learning and statistics for dependent discrete event data. To date, most existing models assume stationary kernels (including the classical Hawkes processes) and simple parametric models. Modern applications with complex event data require more general point process models that can incorporate contextual information of the events, called marks, besides the temporal and location information. Moreover, such applications often require non-stationary models to capture more complex spatio-temporal dependence. To tackle these challenges, a key question is to devise a versatile influence kernel in the point process model. In this paper, we introduce a novel and general neural network-based non-stationary influence kernel with high expressiveness for handling complex discrete events data while providing theoretical performance guarantees. We demonstrate the superior performance of our proposed method compared with the state-of-the-art on synthetic and real data.

How to Inject Backdoors with Better Consistency: Logit Anchoring on Clean Data

- Zhiyuan Zhang, Lingjuan Lyu, Weiqiang Wang, Lichao Sun, Xu Sun
- abstract@open-review(Poster): Since training a large-scale backdoored model from scratch requires a large training dataset, several recent attacks have considered to inject backdoors into a trained clean model without altering model behaviors on the clean data. Previous work finds that backdoors can be injected into a trained clean model with Adversarial Weight Perturbation (AWP), which means the variation of parameters are small in backdoor learning. In this work, we observe an interesting phenomenon that the variations of parameters are always AWPs when tuning the trained clean model to inject backdoors. We further provide theoretical analysis to explain this phenomenon. We are the first to formulate the behavior of maintaining accuracy on clean data as the consistency of backdoored models, which includes both global consistency and instance-wise consistency. We extensively analyze the effects of AWPs on the consistency of backdoored models. In order to achieve better consistency, we propose a novel anchoring loss to anchor or freeze the model behaviors on the clean data, with a theoretical guarantee.

A Biologically Interpretable Graph Convolutional Network to Link Genetic Risk Pathways and Imaging Phenotypes of Disease

- Sayan Ghosal, Qiang Chen, Giulio Pergola, Aaron L Goldman, William Ulrich, Daniel R Weinberger, Archana Venkataraman
- abstract@open-review(Poster): We propose a novel end-to-end framework for whole-brain and whole-genome imaging-genetics. Our genetics network uses hierarchical graph convolution and pooling operations to embed subject-level data onto a low-dimensional latent space. The hierarchical network implicitly tracks the convergence of genetic risk across well-established biological pathways, while an attention mechanism automatically identifies the salient edges of this network at the subject level. In parallel, our imaging network projects multimodal data onto a set of latent embeddings. For interpretability, we implement a Bayesian feature selection strategy to extract the discriminative imaging biomarkers; these feature weights are optimized alongside the other model parameters. We couple the imaging and genetic embeddings with a predictor network, to ensure that the learned representations are linked to phenotype. We evaluate our framework on a schizophrenia dataset that includes two functional MRI paradigms and gene scores derived from Single Nucleotide Polymorphism data. Using repeated 10-fold cross-validation, we show that our imaging-genetics fusion achieves the better classification performance than state-of-the-art baselines. In an exploratory analysis, we further show that the biomarkers identified by our model are reproducible and closely associated with deficits in schizophrenia.

Differentiable Prompt Makes Pre-trained Language Models Better Few-shot Learners

- Ningyu Zhang, Luoqiu Li, Xiang Chen, Shumin Deng, Zhen Bi, Chuanqi Tan, Fei Huang, Huajun Chen
- abstract@open-review(Poster): Large-scale pre-trained language models have contributed significantly to natural language processing by demonstrating remarkable abilities as few-shot learners. However, their effectiveness depends mainly on scaling the model parameters and prompt design, hindering their implementation in most real-world applications. This study proposes a novel pluggable, extensible, and efficient approach named DifferentiAble pRompT (DART), which can convert small language models into better few-shot learners. The main principle behind this approach involves reformulating potential natural language processing tasks into the task of a pre-trained language model and differentially optimizing the prompt template as well as the target label with backpropagation. Furthermore, the proposed approach can be: (i) Plugged to any pre-trained language models; (ii) Extended to widespread classification tasks. A comprehensive evaluation of standard NLP tasks demonstrates that the proposed approach achieves a better few-shot performance.

OntoProtein: Protein Pretraining With Gene Ontology Embedding

- Ningyu Zhang, Zhen Bi, Xiaozhuan Liang, Siyuan Cheng, Haosen Hong, Shumin Deng, Qiang Zhang, Jiazhang Lian, Huajun Chen
- abstract@open-review(Poster): Self-supervised protein language models have proved their effectiveness in learning the proteins representations. With the increasing computational power, current protein language models pre-trained with millions of diverse sequences can advance the parameter scale from million-level to billion-level and achieve remarkable improvement. However, those prevailing approaches rarely consider incorporating knowledge graphs (KGs), which can provide rich structured knowledge facts for better protein representations. We argue that informative biology knowledge in KGs can enhance protein representation with external knowledge. In this work, we propose OntoProtein, the first general framework that makes use of structure in GO (Gene Ontology) into protein pre-training models. We construct a novel large-scale knowledge graph that consists of GO and its related proteins, and gene annotation texts or protein sequences describe all nodes in the graph. We propose novel contrastive learning with knowledge-aware negative sampling to jointly optimize the knowledge graph and protein embedding during pre-training. Experimental results show that OntoProtein can surpass state-of-the-art methods with pre-trained protein language models in TAPE benchmark and yield better performance compared with baselines in protein-protein interaction and protein function prediction.

Permutation Compressors for Provably Faster Distributed Nonconvex Optimization

- RafaÅ, Szlendak, Alexander Tyurin, Peter RichtÃ¡rik
- abstract@open-review(Poster): In this work we study the MARINA method of Gorbunov et al (ICML, 2021) -- the current state-of-the-art distributed non-convex optimization method in terms of theoretical communication complexity. Theoretical superiority of this method can be largely attributed to two sources: a carefully engineered biased stochastic gradient estimator, which leads to a reduction in the number of communication rounds, and the reliance on {\em independent} stochastic communication compression, which leads to a reduction in the number of transmitted bits within each communication round. In this paper we i) extend the theory of MARINA to support a much wider class of potentially {\em correlated} compressors, extending the reach of the method beyond the classical independent compressors setting, ii) show that a new quantity, for which we coin the name {\em Hessian variance}, allows us to significantly refine the original analysis of MARINA without any additional assumptions, and iii) identify a special class of correlated compressors based on the idea of {\em random permutations}, for which we coin the term Perm$K$, the use of which leads to up to $O(\sqrt{n})$ (resp. $O(1 + d/\sqrt{n})$) improvement in the theoretical communication complexity of MARINA in the low Hessian variance regime when $d\geq n$ (resp. $d \leq n$), where n is the number of workers and d is the number of parameters describing the model we are learning. We corroborate our theoretical results with carefully engineered synthetic experiments with minimizing the average of nonconvex quadratics, and on autoencoder training with the MNIST dataset.

Few-shot Learning via Dirichlet Tessellation Ensemble

- Chunwei Ma, Ziyun Huang, Mingchen Gao, Jinhui Xu
- abstract@open-review(Poster): Few-shot learning (FSL) is the process of rapid generalization from abundant base samples to inadequate novel samples. Despite extensive research in recent years, FSL is still not yet able to generate satisfactory solutions for a wide range of real-world applications. To confront this challenge, we study the FSL problem from a geometric point of view in this paper. One observation is that the widely embraced ProtoNet model is essentially a Voronoi Diagram (VD) in the feature space. We retrofit it by making use of a recent advance in computational geometry called Cluster-induced Voronoi Diagram (CIVD). Starting from the simplest nearest neighbor model, CIVD gradually incorporates cluster-to-point and then cluster-to-cluster relationships for space subdivision, which is used to improve the accuracy and robustness at multiple stages of FSL. Specifically, we use CIVD (1) to integrate parametric and nonparametric few-shot classifiers; (2) to combine feature representation and surrogate representation; (3) and to leverage feature-level, transformation-level, and geometry-level heterogeneities for a better ensemble. Our CIVD-based workflow enables us to achieve new state-of-the-art results on mini-ImageNet, CUB, and tiered-ImagenNet datasets, with ${\sim}2\%{-}5\%$ improvements upon the next best. To summarize, CIVD provides a mathematically elegant and geometrically interpretable framework that compensates for extreme data insufficiency, prevents overfitting, and allows for fast geometric ensemble for thousands of individual VD. These together make FSL stronger.

Deep Point Cloud Reconstruction

- Jaesung Choe, ByeongIn Joung, Francois Rameau, Jaesik Park, In So Kweon
- abstract@open-review(Poster): Point cloud obtained from 3D scanning is often sparse, noisy, and irregular. To cope with these issues, recent studies have been separately conducted to densify, denoise, and complete inaccurate point cloud. In this paper, we advocate that jointly solving these tasks leads to significant improvement for point cloud reconstruction. To this end, we propose a deep point cloud reconstruction network consisting of two stages: 1) a 3D sparse stacked-hourglass network as for the initial densification and denoising, 2) a refinement via transformers converting the discrete voxels into continuous 3D points. In particular, we further improve the performance of the transformers by a newly proposed module called amplified positional encoding. This module has been designed to differently amplify the magnitude of positional encoding vectors based on the points' distances for adaptive refinements. Extensive experiments demonstrate that our network achieves state-of-the-art performance among the recent studies in the ScanNet, ICL-NUIM, and ShapeNet datasets. Moreover, we underline the ability of our network to generalize toward real-world and unmet scenes.

β-Intact-VAE: Identifying and Estimating Causal Effects under Limited Overlap

- Pengzhou Abel Wu, Kenji Fukumizu
- abstract@open-review(Poster): As an important problem in causal inference, we discuss the identification and estimation of treatment effects (TEs) under limited overlap; that is, when subjects with certain features belong to a single treatment group. We use a latent variable to model a prognostic score which is widely used in biostatistics and sufficient for TEs; i.e., we build a generative prognostic model. We prove that the latent variable recovers a prognostic score, and the model identifies individualized treatment effects. The model is then learned as β-Intact-VAEâ€“â€“a new type of variational autoencoder (VAE). We derive the TE error bounds that enable representations balanced for treatment groups conditioned on individualized features. The proposed method is compared with recent methods using (semi-)synthetic datasets.

Promoting Saliency From Depth: Deep Unsupervised RGB-D Saliency Detection

- Wei Ji, Jingjing Li, Qi Bi, chuan guo, Jie Liu, Li Cheng

- abstract@open-review(Poster): Growing interests in RGB-D salient object detection (RGB-D SOD) have been witnessed in recent years, owing partly to the popularity of depth sensors and the rapid progress of deep learning techniques. Unfortunately, existing RGB-D SOD methods typically demand large quantity of training images being thoroughly annotated at pixel-level. The laborious and time-consuming manual annotation has become a real bottleneck in various practical scenarios. On the other hand, current unsupervised RGB-D SOD methods still heavily rely on handcrafted feature representations. This inspires us to propose in this paper a deep unsupervised RGB-D saliency detection approach, which requires no manual pixel-level annotation during training. It is realized by two key ingredients in our training pipeline. First, a depth-disentangled saliency update (DSU) framework is designed to automatically produce pseudo-labels with iterative follow-up refinements, which provides more trustworthy supervision signals for training the saliency network. Second, an attentive training strategy is introduced to tackle the issue of noisy pseudo-labels, by properly re-weighting to highlight the more reliable pseudo-labels. Extensive experiments demonstrate the superior efficiency and effectiveness of our approach in tackling the challenging unsupervised RGB-D SOD scenarios. Moreover, our approach can also be adapted to work in fully-supervised situation. Empirical studies show the incorporation of our approach gives rise to notably performance improvement in existing supervised RGB-D SOD models.

Retriever: Learning Content-Style Representation as a Token-Level Bipartite Graph

- Dacheng Yin, Xuanchi Ren, Chong Luo, Yuwang Wang, Zhiwei Xiong, Wenjun Zeng
- abstract@open-review(Poster): This paper addresses the unsupervised learning of content-style decomposed representation. We first give a definition of style and then model the content-style representation as a token-level bipartite graph. An unsupervised framework, named Retriever, is proposed to learn such representations. First, a cross-attention module is employed to retrieve permutation invariant (P.I.) information, defined as style, from the input data. Second, a vector quantization (VQ) module is used, together with man-induced constraints, to produce interpretable content tokens. Last, an innovative link attention module serves as the decoder to reconstruct data from the decomposed content and style, with the help of the linking keys. Being modal-agnostic, the proposed Retriever is evaluated in both speech and image domains. The state-of-the-art zero-shot voice conversion performance confirms the disentangling ability of our framework. Top performance is also achieved in the part discovery task for images, verifying the interpretability of our representation. In addition, the vivid part-based style transfer quality demonstrates the potential of Retriever to support various fascinating generative tasks. Project page at https://ydcustc.github.io/retriever-demo/.

Neural Markov Controlled SDE: Stochastic Optimization for Continuous-Time Data

- Sung Woo Park, Kyungjae Lee, Junseok Kwon
- abstract@open-review(Poster): We propose a novel probabilistic framework for modeling stochastic dynamics with the rigorous use of stochastic optimal control theory. The proposed model called the neural Markov controlled stochastic differential equation (CSDE) overcomes the fundamental and structural limitations of conventional dynamical models by introducing the following two components: (1) Markov dynamic programming to efficiently train the proposed CSDE and (2) multi-conditional forward-backward losses to provide rich information for accurate inference and to assure theoretical optimality. We demonstrate that our dynamical model efficiently generates a complex time series in the data space without extra networks while showing comparable performance against existing model-based methods on several datasets.

CrossFormer: A Versatile Vision Transformer Hinging on Cross-scale Attention

- Wenxiao Wang, Lu Yao, Long Chen, Binbin Lin, Deng Cai, Xiaofei He, Wei Liu
- abstract@open-review(Poster): Transformers have made great progress in dealing with computer vision tasks. However, existing vision transformers have not yet possessed the ability of building the interactions among features of different scales, which is perceptually important to visual inputs. The reasons are two-fold: (1) Input embeddings of each layer are equal-scale, so no cross-scale feature can be extracted; (2) to lower the computational cost, some vision transformers merge adjacent embeddings inside the self-attention module, thus sacrificing small-scale (fine-grained) features of the embeddings and also disabling the cross-scale interactions. To this end, we propose Cross-scale Embedding Layer (CEL) and Long Short Distance Attention (LSDA). On the one hand, CEL blends each embedding with multiple patches of different scales, providing the self-attention module itself with cross-scale features. On the other hand, LSDA splits the self-attention module into a short-distance one and a long-distance counterpart, which not only reduces the computational burden but also keeps both small-scale and large-scale features in the embeddings. Through the above two designs, we achieve cross-scale attention. Besides, we put forward a dynamic position bias for vision transformers to make the popular relative position bias apply to variable-sized images. Hinging on the cross-scale attention module, we construct a versatile vision architecture, dubbed CrossFormer, which accommodates variable-sized inputs. Extensive experiments show that CrossFormer outperforms the other vision transformers on image classification, object detection, instance segmentation, and semantic segmentation tasks.

Adversarially Robust Conformal Prediction

- Asaf Gendler, Tsui-Wei Weng, Luca Daniel, Yaniv Romano
- abstract@open-review(Poster): Conformal prediction is a model-agnostic tool for constructing prediction sets that are valid under the common i.i.d. assumption, which has been applied to quantify the prediction uncertainty of deep net classifiers. In this paper, we generalize this framework to the case where adversaries exist during inference time, under which the i.i.d. assumption is grossly violated. By combining conformal prediction with randomized smoothing, our proposed method forms a prediction set with finite-sample coverage guarantee that holds for any data distribution with ℓ_2-norm bounded adversarial noise, generated by any adversarial attack algorithm. The core idea is to bound the Lipschitz constant of the non-conformity score by smoothing it with Gaussian noise and leverage this knowledge to account for the effect of the unknown adversarial perturbation. We demonstrate the necessity of our method in the adversarial setting and the validity of our theoretical guarantee on three widely used benchmark data sets: CIFAR10, CIFAR100, and ImageNet.

Hot-Refresh Model Upgrades with Regression-Free Compatible Training in Image Retrieval

- Binjie Zhang, Yixiao Ge, Yantao Shen, Yu Li, Chun Yuan, XUYUAN XU, Yexin Wang, Ying Shan
- abstract@open-review(Poster): The task of hot-refresh model upgrades of image retrieval systems plays an essential role in the industry but has never been investigated in academia before. Conventional cold-refresh model upgrades can only deploy new models after the gallery is overall backfilled, taking weeks or even months for massive data. In contrast, hot-refresh model upgrades deploy the new model immediately and then gradually improve the retrieval accuracy by backfilling the gallery on-the-fly. Compatible training has made it possible, however, the problem of model regression with negative flips poses a great challenge to the stable improvement of user experience. We argue that it is mainly due to the fact that new-to-old positive query-gallery pairs may show less similarity than new-to-new negative pairs. To solve the problem, we introduce a Regression-Alleviating Compatible Training (RACT) method to properly constrain the feature compatibility while reducing negative flips. The core is to encourage the new-to-old positive pairs to be more similar than both the new-to-old negative pairs and the new-to-new negative pairs. An efficient uncertainty-based backfilling strategy is further introduced to fasten accuracy improvements. Extensive experiments on large-scale retrieval benchmarks (e.g., Google Landmark) demonstrate that our RACT effectively alleviates the model regression for one more step towards seamless model upgrades.

Visual Representation Learning over Latent Domains

- Lucas Deecke, Timothy Hospedales, Hakan Bilen
- abstract@open-review(Poster): A fundamental shortcoming of deep neural networks is their specialization to a single task and domain. While multi-domain learning enables the learning of compact models that span multiple visual domains, these rely on the presence of domain labels, in turn requiring

laborious curation of datasets. This paper proposes a less explored, but highly realistic new setting called latent domain learning: learning over data from different domains, without access to domain annotations. Experiments show that this setting is challenging for standard models and existing multi-domain approaches, calling for new customized solutions: a sparse adaptation strategy is formulated which enhances performance by accounting for latent domains in data. Our method can be paired seamlessly with existing models, and benefits conceptually related tasks, e.g. empirical fairness problems and long-tailed recognition.

Chemical-Reaction-Aware Molecule Representation Learning

- Hongwei Wang, Weijiang Li, Xiaomeng Jin, Kyunghyun Cho, Heng Ji, Jiawei Han, Martin Burke
- abstract@open-review(Poster): Molecule representation learning (MRL) methods aim to embed molecules into a real vector space. However, existing SMILES-based (Simplified Molecular-Input Line-Entry System) or GNN-based (Graph Neural Networks) MRL methods either take SMILES strings as input that have difficulty in encoding molecule structure information, or over-emphasize the importance of GNN architectures but neglect their generalization ability. Here we propose using chemical reactions to assist learning molecule representation. The key idea of our approach is to preserve the equivalence of molecules with respect to chemical reactions in the embedding space, i.e., forcing the sum of reactant embeddings and the sum of product embeddings to be equal for each chemical equation. This constraint is proven effective to 1) keep the embedding space well-organized and 2) improve the generalization ability of molecule embeddings. Moreover, our model can use any GNN as the molecule encoder and is thus agnostic to GNN architectures. Experimental results demonstrate that our method achieves state-of-the-art performance in a variety of downstream tasks, e.g., reaction product prediction, molecule property prediction, reaction classification, and graph-edit-distance prediction. The code is available at https://github.com/hwwang55/MolR.

Skill-based Meta-Reinforcement Learning

- Taewook Nam, Shao-Hua Sun, Karl Pertsch, Sung Ju Hwang, Joseph J Lim
- abstract@open-review(Poster): While deep reinforcement learning methods have shown impressive results in robot learning, their sample inefficiency makes the learning of complex, long-horizon behaviors with real robot systems infeasible. To mitigate this issue, meta-reinforcement learning methods aim to enable fast learning on novel tasks by learning how to learn. Yet, the application has been limited to short-horizon tasks with dense rewards. To enable learning long-horizon behaviors, recent works have explored leveraging prior experience in the form of offline datasets without reward or task annotations. While these approaches yield improved sample efficiency, millions of interactions with environments are still required to solve complex tasks. In this work, we devise a method that enables meta-learning on long-horizon, sparse-reward tasks, allowing us to solve unseen target tasks with orders of magnitude fewer environment interactions. Our core idea is to leverage prior experience extracted from offline datasets during meta-learning. Specifically, we propose to (1) extract reusable skills and a skill prior from offline datasets, (2) meta-train a high-level policy that learns to efficiently compose learned skills into long-horizon behaviors, and (3) rapidly adapt the meta-trained policy to solve an unseen target task. Experimental results on continuous control tasks in navigation and manipulation demonstrate that the proposed method can efficiently solve long-horizon novel target tasks by combining the strengths of meta-learning and the usage of offline datasets, while prior approaches in RL, meta-RL, and multi-task RL require substantially more environment interactions to solve the tasks.

InfinityGAN: Towards Infinite-Pixel Image Synthesis

- Chieh Hubert Lin, Hsin-Ying Lee, Yen-Chi Cheng, Sergey Tulyakov, Ming-Hsuan Yang
- abstract@open-review(Poster): We present InfinityGAN, a method to generate arbitrary-sized images. The problem is associated with several key challenges. First, scaling existing models to an arbitrarily large image size is resource-constrained, both in terms of computation and availability of large-field-of-view training data. InfinityGAN trains and infers patch-by-patch seamlessly with low computational resources. Second, large images should be locally and globally consistent, avoid repetitive patterns, and look realistic. To address these, InfinityGAN takes global appearance, local structure and texture into account. With this formulation, we can generate images with spatial size and level of detail not attainable before. Experimental evaluation supports that InfinityGAN generates images with superior global structure compared to baselines and features parallelizable inference. Finally, we show several applications unlocked by our approach, such as fusing styles spatially, multi-modal outpainting and image inbetweening at arbitrary input and output sizes.

Shuffle Private Stochastic Convex Optimization

- Albert Cheu, Matthew Joseph, Jieming Mao, Binghui Peng
- abstract@open-review(Poster): In shuffle privacy, each user sends a collection of randomized messages to a trusted shuffler, the shuffler randomly permutes these messages, and the resulting shuffled collection of messages must satisfy differential privacy. Prior work in this model has largely focused on protocols that use a single round of communication to compute algorithmic primitives like means, histograms, and counts. In this work, we present interactive shuffle protocols for stochastic convex optimization. Our optimization protocols rely on a new noninteractive protocol for summing vectors of bounded ℓ_2 norm. By combining this sum subroutine with techniques including mini-batch stochastic gradient descent, accelerated gradient descent, and Nesterov's smoothing method, we obtain loss guarantees for a variety of convex loss functions that significantly improve on those of the local model and sometimes match those of the central model.

Know Your Action Set: Learning Action Relations for Reinforcement Learning

- Ayush Jain, Norio Kosaka, Kyung-Min Kim, Joseph J Lim
- abstract@open-review(Poster): Intelligent agents can solve tasks in various ways depending on their available set of actions. However, conventional reinforcement learning (RL) assumes a fixed action set. This work asserts that tasks with varying action sets require reasoning of the relations between the available actions. For instance, taking a nail-action in a repair task is meaningful only if a hammer-action is also available. To learn and utilize such action relations, we propose a novel policy architecture consisting of a graph attention network over the available actions. We show that our model makes informed action decisions by correctly attending to other related actions in both value-based and policy-based RL. Consequently, it outperforms non-relational architectures on applications where the action space often varies, such as recommender systems and physical reasoning with tools and skills. Results and code at https://sites.google.com/view/varyingaction .

On the Importance of Difficulty Calibration in Membership Inference Attacks

- Lauren Watson, Chuan Guo, Graham Cormode, Alexandre Sablayrolles
- abstract@open-review(Poster): The vulnerability of machine learning models to membership inference attacks has received much attention in recent years. However, existing attacks mostly remain impractical due to having high false positive rates, where non-member samples are often erroneously predicted as members. This type of error makes the predicted membership signal unreliable, especially since most samples are non-members in real world applications. In this work, we argue that membership inference attacks can benefit drastically from difficulty calibration, where an attack's predicted membership score is adjusted to the difficulty of correctly classifying the target sample. We show that difficulty calibration can significantly reduce the false positive rate of a variety of existing attacks without a loss in accuracy.

Entroformer: A Transformer-based Entropy Model for Learned Image Compression

- Yichen Qian, Xiuyu Sun, Ming Lin, Zhiyu Tan, Rong Jin
- abstract@open-review(Poster): One critical component in lossy deep image compression is the entropy model, which predicts the probability distribution of the quantized latent representation in the encoding and decoding modules. Previous works build entropy models upon convolutional neural networks which are inefficient in capturing global dependencies. In this work, we propose a novel transformer-based entropy model, termed Entroformer, to capture long-range dependencies in probability distribution estimation effectively and efficiently. Different from vision transformers in image classification, the Entroformer is highly optimized for image compression, including a top-k self-attention and a diamond relative position encoding. Meanwhile, we further expand this architecture with a parallel bidirectional context model to speed up the decoding process. The experiments show that the Entroformer achieves state-of-the-art performance on image compression while being time-efficient.

Dual Lottery Ticket Hypothesis

- Yue Bai, Huan Wang, ZHIQIANG TAO, Kunpeng Li, Yun Fu
- abstract@open-review(Poster): Fully exploiting the learning capacity of neural networks requires overparameterized dense networks. On the other side, directly training sparse neural networks typically results in unsatisfactory performance. Lottery Ticket Hypothesis (LTH) provides a novel view to investigate sparse network training and maintain its capacity. Concretely, it claims there exist winning tickets from a randomly initialized network found by iterative magnitude pruning and preserving promising trainability (or we say being in trainable condition). In this work, we regard the winning ticket from LTH as the subnetwork which is in trainable condition and its performance as our benchmark, then go from a complementary direction to articulate the Dual Lottery Ticket Hypothesis (DLTH): Randomly selected subnetworks from a randomly initialized dense network can be transformed into a trainable condition and achieve admirable performance compared with LTH --- random tickets in a given lottery pool can be transformed into winning tickets. Specifically, by using uniform-randomly selected subnetworks to represent the general cases, we propose a simple sparse network training strategy, Random Sparse Network Transformation (RST), to substantiate our DLTH. Concretely, we introduce a regularization term to borrow learning capacity and realize information extrusion from the weights which will be masked. After finishing the transformation for the randomly selected subnetworks, we conduct the regular finetuning to evaluate the model using fair comparisons with LTH and other strong baselines. Extensive experiments on several public datasets and comparisons with competitive approaches validate our DLTH as well as the effectiveness of the proposed model RST. Our work is expected to pave a way for inspiring new research directions of sparse network training in the future. Our code is available at https://github.com/yueb17/DLTH.

GNN is a Counter? Revisiting GNN for Question Answering

- Kuan Wang, Yuyu Zhang, Diyi Yang, Le Song, Tao Qin
- abstract@open-review(Poster): Question Answering (QA) has been a long-standing research topic in AI and NLP fields, and a wealth of studies has been conducted to attempt to equip QA systems with human-level reasoning capability. To approximate the complicated human reasoning process, state-of-the-art QA systems commonly use pre-trained language models (LMs) to access knowledge encoded in LMs together with elaborately designed modules based on Graph Neural Networks (GNNs) to perform reasoning over knowledge graphs (KGs). However, many problems remain open regarding the reasoning functionality of these GNN-based modules. Can these GNN-based modules really perform a complex reasoning process? Are they under- or over-complicated for QA? To open the black box of GNN and investigate these problems, we dissect state-of-the-art GNN modules for QA and analyze their reasoning capability. We discover that even a very simple graph neural counter can outperform all the existing GNN modules on CommonsenseQA and OpenBookQA, two popular QA benchmark datasets which heavily rely on knowledge-aware reasoning. Our work reveals that existing knowledge-aware GNN modules may only carry out some simple reasoning such as counting. It remains a challenging open problem to build comprehensive reasoning modules for knowledge-powered QA.

IFR-Explore: Learning Inter-object Functional Relationships in 3D Indoor Scenes

- QI LI, Kaichun Mo, Yanchao Yang, Hang Zhao, Leonidas Guibas
- abstract@open-review(Poster): Building embodied intelligent agents that can interact with 3D indoor environments has received increasing research attention in recent years. While most works focus on single-object or agent-object visual functionality and affordances, our work proposes to study a novel, underexplored, kind of visual relations that is also important to perceive and model -- inter-object functional relationships (e.g., a switch on the wall turns on or off the light, a remote control operates the TV). Humans often spend no effort or only a little to infer these relationships, even when entering a new room, by using our strong prior knowledge (e.g., we know that buttons control electrical devices) or using only a few exploratory interactions in cases of uncertainty (e.g., multiple switches and lights in the same room). In this paper, we take the first step in building AI system learning inter-object functional relationships in 3D indoor environments with key technical contributions of modeling prior knowledge by training over large-scale scenes and designing interactive policies for effectively exploring the training scenes and quickly adapting to novel test scenes. We create a new dataset based on the AI2Thor and PartNet datasets and perform extensive experiments that prove the effectiveness of our proposed method.

VAT-Mart: Learning Visual Action Trajectory Proposals for Manipulating 3D ARTiculated Objects

- Ruihai Wu, Yan Zhao, Kaichun Mo, Zizheng Guo, Yian Wang, Tianhao Wu, Qingnan Fan, Xuelin Chen, Leonidas Guibas, Hao Dong
- abstract@open-review(Poster): Perceiving and manipulating 3D articulated objects (e.g., cabinets, doors) in human environments is an important yet challenging task for future home-assistant robots. The space of 3D articulated objects is exceptionally rich in their myriad semantic categories, diverse shape geometry, and complicated part functionality. Previous works mostly abstract kinematic structure with estimated joint parameters and part poses as the visual representations for manipulating 3D articulated objects. In this paper, we propose object-centric actionable visual priors as a novel perception-interaction handshaking point that the perception system outputs more actionable guidance than kinematic structure estimation, by predicting dense geometry-aware, interaction-aware, and task-aware visual action affordance and trajectory proposals. We design an interaction-for-perception framework VAT-Mart to learn such actionable visual representations by simultaneously training a curiosity-driven reinforcement learning policy exploring diverse interaction trajectories and a perception module summarizing and generalizing the explored knowledge for pointwise predictions among diverse shapes. Experiments prove the effectiveness of the proposed approach using the large-scale PartNet-Mobility dataset in SAPIEN environment and show promising generalization capabilities to novel test shapes, unseen object categories, and real-world data.

Neural graphical modelling in continuous-time: consistency guarantees and algorithms

- Alexis Bellot, Kim Branson, Mihaela van der Schaar
- abstract@open-review(Poster): The discovery of structure from time series data is a key problem in fields of study working with complex systems. Most identifiability results and learning algorithms assume the underlying dynamics to be discrete in time. Comparatively few, in contrast, explicitly define dependencies in infinitesimal intervals of time, independently of the scale of observation and of the regularity of sampling. In this paper, we consider score-based structure learning for the study of dynamical systems. We prove that for vector fields parameterized in a large class of neural networks, least squares optimization with adaptive regularization schemes consistently recovers directed graphs of local independencies in systems of stochastic differential equations. Using this insight, we propose a score-based learning algorithm based on penalized Neural Ordinary Differential Equations (modelling the mean process) that we show to be applicable to the general setting of irregularly-sampled multivariate time series and to outperform the state of the art across a range of dynamical systems.

C-Planning: An Automatic Curriculum for Learning Goal-Reaching Tasks

- Tianjun Zhang, Benjamin Eysenbach, Ruslan Salakhutdinov, Sergey Levine, Joseph E. Gonzalez

- abstract@open-review(Poster): Goal-conditioned reinforcement learning (RL) has shown great success recently at solving a wide range of tasks(e.g., navigation, robotic manipulation). However, learning to reach distant goals remains a central challenge to the field, and the task is particularly hard without any offline data, expert demonstrations, and reward shaping. In this paper, we propose to solve the distant goal-reaching task by using search at training time to generate a curriculum of intermediate states. Specifically, we introduce the algorithm Classifier-Planning (C-Planning) by framing the learning of the goal-conditioned policies as variational inference. C-Planning naturally follows expectation maximization (EM): the E-step corresponds to planning an optimal sequence of waypoints using graph search, while the M-step aims to learn a goal-conditioned policy to reach those waypoints. One essential difficulty of designing such an algorithm is accurately modeling the distribution over way-points to sample from. In C-Planning, we propose to sample the waypoints using contrastive methods to learn a value function. Unlike prior methods that combine goal-conditioned RL with graph search, ours performs search only during training and not testing, significantly decreasing the compute costs of deploying the learned policy. Empirically, we demonstrate that our method not only improves the sample efficiency of prior methods but also successfully solves temporally extended navigation and manipulation tasks, where prior goal-conditioned RL methods (including those based on graph search) fail to solve.

NAS-Bench-Suite: NAS Evaluation is (Now) Surprisingly Easy

- Yash Mehta, Colin White, Arber Zela, Arjun Krishnakumar, Guri Zabergja, Shakiba Moradian, Mahmoud Safari, Kaicheng Yu, Frank Hutter
- abstract@open-review(Poster): The release of tabular benchmarks, such as NAS-Bench-101 and NAS-Bench-201, has significantly lowered the computational overhead for conducting scientific research in neural architecture search (NAS). Although they have been widely adopted and used to tune real-world NAS algorithms, these benchmarks are limited to small search spaces and focus solely on image classification. Recently, several new NAS benchmarks have been introduced that cover significantly larger search spaces over a wide range of tasks, including object detection, speech recognition, and natural language processing. However, substantial differences among these NAS benchmarks have so far prevented their widespread adoption, limiting researchers to using just a few benchmarks. In this work, we present an in-depth analysis of popular NAS algorithms and performance prediction methods across 25 different combinations of search spaces and datasets, finding that many conclusions drawn from a few NAS benchmarks do \emph{not} generalize to other benchmarks. To help remedy this problem, we introduce \nasbs, a comprehensive and extensible collection of NAS benchmarks, accessible through a unified interface, created with the aim to facilitate reproducible, generalizable, and rapid NAS research. Our code is available at https://github.com/automl/naslib.

Machine Learning For Elliptic PDEs: Fast Rate Generalization Bound, Neural Scaling Law and Minimax Optimality

- Yiping Lu, Haoxuan Chen, Jianfeng Lu, Lexing Ying, Jose Blanchet
- abstract@open-review(Poster): In this paper, we study the statistical limits of deep learning techniques for solving elliptic partial differential equations (PDEs) from random samples using the Deep Ritz Method (DRM) and Physics-Informed Neural Networks (PINNs). To simplify the problem, we focus on a prototype elliptic PDE: the Schr\"odinger equation on a hypercube with zero Dirichlet boundary condition, which has wide application in the quantum-mechanical systems. We establish upper and lower bounds for both methods, which improves upon concurrently developed upper bounds for this problem via a fast rate generalization bound. We discover that the current Deep Ritz Methods is sub-optimal and propose a modified version of it. We also prove that PINN and the modified version of DRM can achieve minimax optimal bounds over Sobolev spaces. Empirically, following recent work which has shown that the deep model accuracy will improve with growing training sets according to a power law, we supply computational experiments to show a similar behavior of dimension dependent power law for deep PDE solvers.

Variational oracle guiding for reinforcement learning

- Dongqi Han, Tadashi Kozuno, Xufang Luo, Zhao-Yun Chen, Kenji Doya, Yuqing Yang, Dongsheng Li
- abstract@open-review(Poster): How to make intelligent decisions is a central problem in machine learning and artificial intelligence. Despite recent successes of deep reinforcement learning (RL) in various decision making problems, an important but under-explored aspect is how to leverage oracle observation (the information that is invisible during online decision making, but is available during offline training) to facilitate learning. For example, human experts will look at the replay after a Poker game, in which they can check the opponents' hands to improve their estimation of the opponents' hands from the visible information during playing. In this work, we study such problems based on Bayesian theory and derive an objective to leverage oracle observation in RL using variational methods. Our key contribution is to propose a general learning framework referred to as variational latent oracle guiding (VLOG) for DRL. VLOG is featured with preferable properties such as its robust and promising performance and its versatility to incorporate with any value-based DRL algorithm. We empirically demonstrate the effectiveness of VLOG in online and offline RL domains with tasks ranging from video games to a challenging tile-based game Mahjong. Furthermore, we publish the Mahjong environment and an offline RL dataset as a benchmark to facilitate future research on oracle guiding (https://github.com/Agony5757/mahjong).

CDTrans: Cross-domain Transformer for Unsupervised Domain Adaptation

- Tongkun Xu, Weihua Chen, Pichao WANG, Fan Wang, Hao Li, Rong Jin
- abstract@open-review(Poster): Unsupervised domain adaptation (UDA) aims to transfer knowledge learned from a labeled source domain to a different unlabeled target domain. Most existing UDA methods focus on learning domain-invariant feature representation, either from the domain level or category level, using convolution neural networks (CNNs)-based frameworks. One fundamental problem for the category level based UDA is the production of pseudo labels for samples in target domain, which are usually too noisy for accurate domain alignment, inevitably compromising the UDA performance. With the success of Transformer in various tasks, we find that the cross-attention in Transformer is robust to the noisy input pairs for better feature alignment, thus in this paper Transformer is adopted for the challenging UDA task. Specifically, to generate accurate input pairs, we design a two-way center-aware labeling algorithm to produce pseudo labels for target samples. Along with the pseudo labels, a weight-sharing triple-branch transformer framework is proposed to apply self-attention and cross-attention for source/target feature learning and source-target domain alignment, respectively. Such design explicitly enforces the framework to learn discriminative domain-specific and domain-invariant representations simultaneously. The proposed method is dubbed CDTrans (cross-domain transformer), and it provides one of the first attempts to solve UDA tasks with a pure transformer solution. Experiments show that our proposed method achieves the best performance on public UDA datasets, e.g. VisDA-2017 and DomainNet. Code and models are available at https://github.com/CDTrans/CDTrans.

Beyond ImageNet Attack: Towards Crafting Adversarial Examples for Black-box Domains

- Qilong Zhang, Xiaodan Li, YueFeng Chen, Jingkuan Song, Lianli Gao, Yuan He, Hui Xue'
- abstract@open-review(Poster): Adversarial examples have posed a severe threat to deep neural networks due to their transferable nature. Currently, various works have paid great efforts to enhance the cross-model transferability, which mostly assume the substitute model is trained in the same domain as the target model. However, in reality, the relevant information of the deployed model is unlikely to leak. Hence, it is vital to build a more practical black-box threat model to overcome this limitation and evaluate the vulnerability of deployed models. In this paper, with only the knowledge of the ImageNet domain, we propose a Beyond ImageNet Attack (BIA) to investigate the transferability towards black-box domains (unknown classification tasks). Specifically, we leverage a generative model to learn the adversarial function for disrupting low-level features of input images. Based on this framework, we further propose two variants to narrow the gap between the source and target domains from the data and model perspectives, respectively. Extensive experiments on coarse-grained and fine-grained domains demonstrate the effectiveness of our proposed methods. Notably, our methods outperform state-of-the-art approaches by up to 7.71\% (towards coarse-grained domains) and 25.91\% (towards fine-grained domains) on average. Our code is available at \url{https://github.com/Alibaba-AAIG/Beyond-ImageNet-Attack}.

Learning to Schedule Learning rate with Graph Neural Networks

- Yuanhao Xiong, Li-Cheng Lan, Xiangning Chen, Ruochen Wang, Cho-Jui Hsieh
- abstract@open-review(Poster): Recent decades have witnessed great development of stochastic optimization in training deep neural networks. Learning rate scheduling is one of the most important factors that influence the performance of stochastic optimizers like Adam. Traditional methods seek to find a relatively proper scheduling among a limited number of pre-defined rules and might not accommodate a particular target problem. Instead, we propose a novel Graph-Network-based Scheduler (GNS), aiming at learning a specific scheduling mechanism without restrictions to existing principles. By constructing a directed graph for the underlying neural network of the target problem, GNS encodes current dynamics with a graph message passing network and trains an agent to control the learning rate accordingly via reinforcement learning. The proposed scheduler can capture the intermediate layer information while being able to generalize to problems of varying scales. Besides, an efficient reward collection procedure is leveraged to speed up training. We evaluate our framework on benchmarking datasets, Fashion-MNIST and CIFAR10 for image classification, and GLUE for language understanding. GNS shows consistent improvement over popular baselines when training CNN and Transformer models. Moreover, GNS demonstrates great generalization to different datasets and network structures.

SketchODE: Learning neural sketch representation in continuous time

- Ayan Das, Yongxin Yang, Timothy Hospedales, Tao Xiang, Yi-Zhe Song
- abstract@open-review(Poster): Learning meaningful representations for chirographic drawing data such as sketches, handwriting, and flowcharts is a gateway for understanding and emulating human creative expression. Despite being inherently continuous-time data, existing works have treated these as discrete-time sequences, disregarding their true nature. In this work, we model such data as continuous-time functions and learn compact representations by virtue of Neural Ordinary Differential Equations. To this end, we introduce the first continuous-time Seq2Seq model and demonstrate some remarkable properties that set it apart from traditional discrete-time analogues. We also provide solutions for some practical challenges for such models, including introducing a family of parameterized ODE dynamics & continuous-time data augmentation particularly suitable for the task. Our models are validated on several datasets including VectorMNIST, DiDi and Quick, Draw!.

Measuring the Interpretability of Unsupervised Representations via Quantized Reversed Probing

- Iro Laina, Yuki M Asano, Andrea Vedaldi
- abstract@open-review(Poster): Self-supervised visual representation learning has recently attracted significant research interest. While a common way to evaluate self-supervised representations is through transfer to various downstream tasks, we instead investigate the problem of measuring their interpretability, i.e. understanding the semantics encoded in raw representations. We formulate the latter as estimating the mutual information between the representation and a space of manually labelled concepts. To quantify this we introduce a decoding bottleneck: information must be captured by simple predictors, mapping concepts to clusters in representation space. This approach, which we call reverse linear probing, provides a single number sensitive to the semanticity of the representation. This measure is also able to detect when the representation contains combinations of concepts (e.g., "red apple") instead of just individual attributes ("red" and "apple" independently). Finally, we propose to use supervised classifiers to automatically label large datasets in order to enrich the space of concepts used for probing. We use our method to evaluate a large number of self-supervised representations, ranking them by interpretability, highlight the differences that emerge compared to the standard evaluation with linear probes and discuss several qualitative insights. Code at: https://github.com/iro-cp/ssl-qrp.

GradMax: Growing Neural Networks using Gradient Information

- Utku Evci, Bart van Merrienboer, Thomas Unterthiner, Fabian Pedregosa, Max Vladymyrov
- abstract@open-review(Poster): The architecture and the parameters of neural networks are often optimized independently, which requires costly retraining of the parameters whenever the architecture is modified. In this work we instead focus on growing the architecture without requiring costly retraining. We present a method that adds new neurons during training without impacting what is already learned, while improving the training dynamics. We achieve the latter by maximizing the gradients of the new weights and efficiently find the optimal initialization by means of the singular value decomposition (SVD). We call this technique Gradient Maximizing Growth (GradMax) and demonstrate its effectiveness in variety of vision tasks and architectures. We open sourced our code at https://github.com/google-research/growneuron

Online Coreset Selection for Rehearsal-based Continual Learning

- Jaehong Yoon, Divyam Madaan, Eunho Yang, Sung Ju Hwang
- abstract@open-review(Poster): A dataset is a shred of crucial evidence to describe a task. However, each data point in the dataset does not have the same potential, as some of the data points can be more representative or informative than others. This unequal importance among the data points may have a large impact in rehearsal-based continual learning, where we store a subset of the training examples (coreset) to be replayed later to alleviate catastrophic forgetting. In continual learning, the quality of the samples stored in the coreset directly affects the model's effectiveness and efficiency. The coreset selection problem becomes even more important under realistic settings, such as imbalanced continual learning or noisy data scenarios. To tackle this problem, we propose Online Coreset Selection (OCS), a simple yet effective method that selects the most representative and informative coreset at each iteration and trains them in an online manner. Our proposed method maximizes the model's adaptation to a target dataset while selecting high-affinity samples to past tasks, which directly inhibits catastrophic forgetting. We validate the effectiveness of our coreset selection mechanism over various standard, imbalanced, and noisy datasets against strong continual learning baselines, demonstrating that it improves task adaptation and prevents catastrophic forgetting in a sample-efficient manner.

Switch to Generalize: Domain-Switch Learning for Cross-Domain Few-Shot Classification

- Zhengdong Hu, Yifan Sun, Yi Yang
- abstract@open-review(Poster): This paper considers few-shot learning under the cross-domain scenario. The cross-domain setting imposes a critical challenge, i.e., using very few (support) samples to generalize the already-learned model to a novel domain. We hold a hypothesis, i.e., if a deep model is capable to fast generalize itself to different domains (using very few samples) during training, it will maintain such domain generalization capacity for testing. It motivates us to propose a novel Domain-Switch Learning (DSL) framework. DSL embeds the cross-domain scenario into the training stage in a ``fast switching'' manner. Specifically, DSL uses a single domain for a training iteration and switches into another domain for the following iteration. During the switching, DSL enforces two constraints: 1) the deep model should not over-fit the domain in the current iteration and 2) the deep model should not forget the already-learned knowledge of other domains. These two constraints jointly promote fast generalization across different domains. Experimental results confirm that the cross-domain generalization capacity can be inherited from the training stage to the testing stage, validating our key hypothesis. Consequentially, DSL significantly improves cross-domain few-shot classification and sets up new state of the art.

Zero-CL: Instance and Feature decorrelation for negative-free symmetric contrastive learning

- Shaofeng Zhang, Feng Zhu, Junchi Yan, Rui Zhao, Xiaokang Yang
- abstract@open-review(Poster): For self-supervised contrastive learning, models can easily collapse and generate trivial constant solutions. The issue has been mitigated by recent improvement on objective design, which however often requires square complexity either for the size of instances ($\mathcal{O}

(N^{2})$) or feature dimensions ($\mathcal{O}(d)^2$). To prevent such collapse, we develop two novel methods by decorrelating on different dimensions on the instance embedding stacking matrix, i.e., \textbf{I}nstance-wise (ICL) and \textbf{F}eature-wise (FCL) \textbf{C}ontrastive \textbf{L}earning. The proposed two methods (FCL, ICL) can be combined synthetically, called Zero-CL, where ``Zero'' means negative samples are \textbf{zero} relevant, which allows Zero-CL to completely discard negative pairs i.e., with \textbf{zero} negative samples. Compared with previous methods, Zero-CL mainly enjoys three advantages: 1) Negative free in symmetric architecture. 2) By whitening transformation, the correlation of the different features is equal to zero, alleviating information redundancy. 3) Zero-CL remains original information to a great extent after transformation, which improves the accuracy against other whitening transformation techniques. Extensive experimental results on CIFAR-10/100 and ImageNet show that Zero-CL outperforms or is on par with state-of-the-art symmetric contrastive learning methods.

Random matrices in service of ML footprint: ternary random features with no performance loss

- Hafiz Tiomoko Ali, Zhenyu Liao, Romain Couillet
- abstract@open-review(Poster): In this article, we investigate the spectral behavior of random features kernel matrices of the type ${\bf K} = \mathbb{E}*{{\bf w}} \left[\sigma\left({\bf w}^{\sf T}{\bf x}_i\right)\sigma\left({\bf w}^{\sf T}{\bf x}_j\right)\right]*{i,j=1}^n$, with nonlinear function $\sigma(\cdot)$, data ${\bf x}*1, \ldots, {\bf x}_n \in \mathbb{R}^p$, and random projection vector ${\bf w} \in \mathbb{R}^p$ having i.i.d. entries. In a high-dimensional setting where the number of data n and their dimension p are both large and comparable, we show, under a Gaussian mixture model for the data, that the eigenspectrum of ${\bf K}$ is independent of the distribution of the i.i.d.(zero-mean and unit-variance) entries of ${\bf w}$, and only depends on $\sigma(\cdot)$ via its (generalized) Gaussian moments $\mathbb{E}{z\sim \mathcal N(0,1)}[\sigma'(z)]$ and $\mathbb{E}_{z\sim \mathcal N(0,1)}[\sigma''(z)]$. As a result, for any kernel matrix ${\bf K}$ of the form above, we propose a novel random features technique, called Ternary Random Features (TRFs), that (i) asymptotically yields the same limiting kernel as the original ${\bf K}$ in a spectral sense and (ii) can be computed and stored much more efficiently, by wisely tuning (in a data-dependent manner) the function σ and the random vector ${\bf w}$, both taking values in ${-1,0,1}$. The computation of the proposed random features requires no multiplication, and a factor of b times less bits for storage compared to classical random features such as random Fourier features, with b the number of bits to store full precision values. Besides, it appears in our experiments on real data that the substantial gains in computation and storage are accompanied with somewhat improved performances compared to state-of-the-art random features methods.

Global Convergence of Multi-Agent Policy Gradient in Markov Potential Games

- Stefanos Leonardos, Will Overman, Ioannis Panageas, Georgios Piliouras
- abstract@open-review(Poster): Potential games are arguably one of the most important and widely studied classes of normal form games. They define the archetypal setting of multi-agent coordination in which all agents utilities are perfectly aligned via a common potential function. Can this intuitive framework be transplanted in the setting of Markov games? What are the similarities and differences between multi-agent coordination with and without state dependence? To answer these questions, we study a natural class of Markov Potential Games (MPGs) that generalize prior attempts at capturing complex stateful multi-agent coordination. Counter-intuitively, insights from normal-form potential games do not carry over as MPGs involve settings where state-games can be zero-sum games. In the opposite direction, Markov games where every state-game is a potential game are not necessarily MPGs. Nevertheless, MPGs showcase standard desirable properties such as the existence of deterministic Nash policies. In our main technical result, we prove convergence of independent policy gradient and its stochastic counterpart to Nash policies (polynomially fast in the approximation error) by adapting recent gradient dominance property arguments developed for single-agent Markov decision processes to multi-agent learning settings.

Rethinking Adversarial Transferability from a Data Distribution Perspective

- Yao Zhu, Jiacheng Sun, Zhenguo Li
- abstract@open-review(Poster): Adversarial transferability enables attackers to generate adversarial examples from the source model to attack the target model, which has raised security concerns about the deployment of DNNs in practice. In this paper, we rethink adversarial transferability from a data distribution perspective and further enhance transferability by score matching based optimization. We identify that some samples with injecting small Gaussian noise can fool different target models, and their adversarial examples under different source models have much stronger transferability. We hypothesize that these samples are in the low-density region of the ground truth distribution where models are not well trained. To improve the attack success rate of adversarial examples, we match the adversarial attacks with the directions which effectively decrease the ground truth density. We propose Intrinsic Adversarial Attack (IAA), which smooths the activation function and decreases the impact of the later layers of a given normal model, to increase the alignment of adversarial attack and the gradient of joint data distribution. We conduct comprehensive transferable attacks against multiple DNNs and show that our IAA can boost the transferability of the crafted attacks in all cases and go beyond state-of-the-art methods.

Transformers Can Do Bayesian Inference

- Samuel MÃ¼ller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, Frank Hutter
- abstract@open-review(Poster): Currently, it is hard to reap the benefits of deep learning for Bayesian methods, which allow the explicit specification of prior knowledge and accurately capture model uncertainty. We present Prior-Data Fitted Networks (PFNs). PFNs leverage large-scale machine learning techniques to approximate a large set of posteriors. The only requirement for PFNs to work is the ability to sample from a prior distribution over supervised learning tasks (or functions). Our method restates the objective of posterior approximation as a supervised classification problem with a set-valued input: it repeatedly draws a task (or function) from the prior, draws a set of data points and their labels from it, masks one of the labels and learns to make probabilistic predictions for it based on the set-valued input of the rest of the data points. Presented with a set of samples from a new supervised learning task as input, PFNs make probabilistic predictions for arbitrary other data points in a single forward propagation, having learned to approximate Bayesian inference. We demonstrate that PFNs can near-perfectly mimic Gaussian processes and also enable efficient Bayesian inference for intractable problems, with over 200-fold speedups in multiple setups compared to current methods. We obtain strong results in very diverse areas such as Gaussian process regression, Bayesian neural networks, classification for small tabular data sets, and few-shot image classification, demonstrating the generality of PFNs. Code and trained PFNs are released at https://github.com/automl/TransformersCanDoBayesianInference.

Learning Discrete Structured Variational Auto-Encoder using Natural Evolution Strategies

- Alon Berliner, Guy Rotman, Yossi Adi, Roi Reichart, Tamir Hazan
- abstract@open-review(Poster): Discrete variational auto-encoders (VAEs) are able to represent semantic latent spaces in generative learning. In many real-life settings, the discrete latent space consists of high-dimensional structures, and propagating gradients through the relevant structures often requires enumerating over an exponentially large latent space. Recently, various approaches were devised to propagate approximated gradients without enumerating over the space of possible structures. In this work, we use Natural Evolution Strategies (NES), a class of gradient-free black-box optimization algorithms, to learn discrete structured VAEs. The NES algorithms are computationally appealing as they estimate gradients with forward pass evaluations only, thus they do not require to propagate gradients through their discrete structures. We demonstrate empirically that optimizing discrete structured VAEs using NES is as effective as gradient-based approximations. Lastly, we prove NES converges for non-Lipschitz functions as appear in discrete structured VAEs.

Learning Features with Parameter-Free Layers

- Dongyoon Han, YoungJoon Yoo, Beomyoung Kim, Byeongho Heo

- abstract@open-review(Poster): Trainable layers such as convolutional building blocks are the standard network design choices by learning parameters to capture the global context through successive spatial operations. When designing an efficient network, trainable layers such as the depthwise convolution is the source of efficiency in the number of parameters and FLOPs, but there was little improvement to the model speed in practice. This paper argues that simple built-in parameter-free operations can be a favorable alternative to the efficient trainable layers replacing spatial operations in a network architecture. We aim to break the stereotype of organizing the spatial operations of building blocks into trainable layers. Extensive experimental analyses based on layer-level studies with fully-trained models and neural architecture searches are provided to investigate whether parameter-free operations such as the max-pool are functional. The studies eventually give us a simple yet effective idea for redesigning network architectures, where the parameter-free operations are heavily used as the main building block without sacrificing the model accuracy as much. Experimental results on the ImageNet dataset demonstrate that the network architectures with parameter-free operations could enjoy the advantages of further efficiency in terms of model speed, the number of the parameters, and FLOPs. Code and ImageNet pretrained models are available at https://github.com/naver-ai/PfLayer.

Denoising Likelihood Score Matching for Conditional Score-based Data Generation

- Chen-Hao Chao, Wei-Fang Sun, Bo-Wun Cheng, Yi-Chen Lo, Chia-Che Chang, Yu-Lun Liu, Yu-Lin Chang, Chia-Ping Chen, Chun-Yi Lee
- abstract@open-review(Poster): Many existing conditional score-based data generation methods utilize Bayes' theorem to decompose the gradients of a log posterior density into a mixture of scores. These methods facilitate the training procedure of conditional score models, as a mixture of scores can be separately estimated using a score model and a classifier. However, our analysis indicates that the training objectives for the classifier in these methods may lead to a serious score mismatch issue, which corresponds to the situation that the estimated scores deviate from the true ones. Such an issue causes the samples to be misled by the deviated scores during the diffusion process, resulting in a degraded sampling quality. To resolve it, we theoretically formulate a novel training objective, called Denoising Likelihood Score Matching (DLSM) loss, for the classifier to match the gradients of the true log likelihood density. Our experimental evidences show that the proposed method outperforms the previous methods on both Cifar-10 and Cifar-100 benchmarks noticeably in terms of several key evaluation metrics. We thus conclude that, by adopting DLSM, the conditional scores can be accurately modeled, and the effect of the score mismatch issue is alleviated.

Memory Replay with Data Compression for Continual Learning

- Liyuan Wang, Xingxing Zhang, Kuo Yang, Longhui Yu, Chongxuan Li, Lanqing HONG, Shifeng Zhang, Zhenguo Li, Yi Zhong, Jun Zhu
- abstract@open-review(Poster): Continual learning needs to overcome catastrophic forgetting of the past. Memory replay of representative old training samples has been shown as an effective solution, and achieves the state-of-the-art (SOTA) performance. However, existing work is mainly built on a small memory buffer containing a few original data, which cannot fully characterize the old data distribution. In this work, we propose memory replay with data compression to reduce the storage cost of old training samples and thus increase their amount that can be stored in the memory buffer. Observing that the trade-off between the quality and quantity of compressed data is highly nontrivial for the efficacy of memory replay, we propose a novel method based on determinantal point processes (DPPs) to efficiently determine an appropriate compression quality for currently-arrived training samples. In this way, using a naive data compression algorithm with a properly selected quality can largely boost recent strong baselines by saving more compressed data in a limited storage space. We extensively validate this across several benchmarks of class-incremental learning and in a realistic scenario of object detection for autonomous driving.

MAML is a Noisy Contrastive Learner in Classification

- Chia Hsiang Kao, Wei-Chen Chiu, Pin-Yu Chen
- abstract@open-review(Poster): Model-agnostic meta-learning (MAML) is one of the most popular and widely adopted meta-learning algorithms, achieving remarkable success in various learning problems. Yet, with the unique design of nested inner-loop and outer-loop updates, which govern the task-specific and meta-model-centric learning, respectively, the underlying learning objective of MAML remains implicit, impeding a more straightforward understanding of it. In this paper, we provide a new perspective of the working mechanism of MAML. We discover that MAML is analogous to a meta-learner using a supervised contrastive objective in classification. The query features are pulled towards the support features of the same class and against those of different classes. Such contrastiveness is experimentally verified via an analysis based on the cosine similarity. Moreover, we reveal that vanilla MAML has an undesirable interference term originating from the random initialization and the cross-task interaction. We thus propose a simple but effective technique, the zeroing trick, to alleviate the interference. Extensive experiments are conducted on both mini-ImageNet and Omniglot datasets to validate the consistent improvement brought by our proposed method.

RelViT: Concept-guided Vision Transformer for Visual Relational Reasoning

- Xiaojian Ma, Weili Nie, Zhiding Yu, Huaizu Jiang, Chaowei Xiao, Yuke Zhu, Song-Chun Zhu, Anima Anandkumar
- abstract@open-review(Poster): Reasoning about visual relationships is central to how humans interpret the visual world. This task remains challenging for current deep learning algorithms since it requires addressing three key technical problems jointly: 1) identifying object entities and their properties, 2) inferring semantic relations between pairs of entities, and 3) generalizing to novel object-relation combinations, i.e., systematic generalization. In this work, we use vision transformers (ViTs) as our base model for visual reasoning and make better use of concepts defined as object entities and their relations to improve the reasoning ability of ViTs. Specifically, we introduce a novel concept-feature dictionary to allow flexible image feature retrieval at training time with concept keys. This dictionary enables two new concept-guided auxiliary tasks: 1) a global task for promoting relational reasoning, and 2) a local task for facilitating semantic object-centric correspondence learning. To examine the systematic generalization of visual reasoning models, we introduce systematic splits for the standard HICO and GQA benchmarks. We show the resulting model, Concept-guided Vision Transformer (or RelViT for short) significantly outperforms prior approaches on HICO and GQA by 16% and 13% in the original split, and by 43% and 18% in the systematic split. Our ablation analyses also reveal our model's compatibility with multiple ViT variants and robustness to hyper-parameters.

Boosted Curriculum Reinforcement Learning

- Pascal Klink, Carlo D'Eramo, Jan Peters, Joni Pajarinen
- abstract@open-review(Poster): Curriculum value-based reinforcement learning (RL) solves a complex target task by reusing action-values across a tailored sequence of related tasks of increasing difficulty. However, finding an exact way of reusing action-values in this setting is still a poorly understood problem. In this paper, we introduce the concept of boosting to curriculum value-based RL, by approximating the action-value function as a sum of residuals trained on each task. This approach, which we refer to as boosted curriculum reinforcement learning (BCRL), has the benefit of naturally increasing the representativeness of the functional space by adding a new residual each time a new task is presented. This procedure allows reusing previous action-values while promoting expressiveness of the action-value function. We theoretically study BCRL as an approximate value iteration algorithm, discussing advantages over regular curriculum RL in terms of approximation accuracy and convergence to the optimal action-value function. Finally, we provide detailed empirical evidence of the benefits of BCRL in problems requiring curricula for accurate action-value estimation and targeted exploration.

ViDT: An Efficient and Effective Fully Transformer-based Object Detector

- Hwanjun Song, Deqing Sun, Sanghyuk Chun, Varun Jampani, Dongyoon Han, Byeongho Heo, Wonjae Kim, Ming-Hsuan Yang
- abstract@open-review(Poster): Transformers are transforming the landscape of computer vision, especially for recognition tasks. Detection transformers are the first fully end-to-end learning systems for object detection, while vision transformers are the first fully transformer-based architecture for image

classification. In this paper, we integrate Vision and Detection Transformers (ViDT) to build an effective and efficient object detector. ViDT introduces a reconfigured attention module to extend the recent Swin Transformer to be a standalone object detector, followed by a computationally efficient transformer decoder that exploits multi-scale features and auxiliary techniques essential to boost the detection performance without much increase in computational load. Extensive evaluation results on the Microsoft COCO benchmark dataset demonstrate that ViDT obtains the best AP and latency trade-off among existing fully transformer-based object detectors, and achieves 49.2AP owing to its high scalability for large models. We release the code and trained models at https://github.com/naver-ai/vidt.

BiBERT: Accurate Fully Binarized BERT

- Haotong Qin, Yifu Ding, Mingyuan Zhang, Qinghua YAN, Aishan Liu, Qingqing Dang, Ziwei Liu, Xianglong Liu
- abstract@open-review(Poster): The large pre-trained BERT has achieved remarkable performance on Natural Language Processing (NLP) tasks but is also computation and memory expensive. As one of the powerful compression approaches, binarization extremely reduces the computation and memory consumption by utilizing 1-bit parameters and bitwise operations. Unfortunately, the full binarization of BERT (i.e., 1-bit weight, embedding, and activation) usually suffer a significant performance drop, and there is rare study addressing this problem. In this paper, with the theoretical justification and empirical analysis, we identify that the severe performance drop can be mainly attributed to the information degradation and optimization direction mismatch respectively in the forward and backward propagation, and propose BiBERT, an accurate fully binarized BERT, to eliminate the performance bottlenecks. Specifically, BiBERT introduces an efficient Bi-Attention structure for maximizing representation information statistically and a Direction-Matching Distillation (DMD) scheme to optimize the full binarized BERT accurately. Extensive experiments show that BiBERT outperforms both the straightforward baseline and existing state-of-the-art quantized BERTs with ultra-low bit activations by convincing margins on the NLP benchmark. As the first fully binarized BERT, our method yields impressive 56.3 times and 31.2 times saving on FLOPs and model size, demonstrating the vast advantages and potential of the fully binarized BERT model in real-world resource-constrained scenarios.

Feature Kernel Distillation

- Bobby He, Mete Ozay
- abstract@open-review(Poster): Trained Neural Networks (NNs) can be viewed as data-dependent kernel machines, with predictions determined by the inner product of last-layer representations across inputs, referred to as the feature kernel. We explore the relevance of the feature kernel for Knowledge Distillation (KD), using a mechanistic understanding of an NNâ€™s optimisation process. We extend the theoretical analysis of Allen-Zhu & Li (2020) to show that a trained NNâ€™s feature kernel is highly dependent on its parameter initialisation, which biases different initialisations of the same architecture to learn different data attributes in a multi-view data setting. This enables us to prove that KD using only pairwise feature kernel comparisons can improve NN test accuracy in such settings, with both single & ensemble teacher models, whereas standard training without KD fails to generalise. We further use our theory to motivate practical considerations for improving student generalisation when using distillation with feature kernels, which allows us to propose a novel approach: Feature Kernel Distillation (FKD). Finally, we experimentally corroborate our theory in the image classification setting, showing that FKD is amenable to ensemble distillation, can transfer knowledge across datasets, and outperforms both vanilla KD & other feature kernel based KD baselines across a range of standard architectures & datasets.

Representation-Agnostic Shape Fields

- Xiaoyang Huang, Jiancheng Yang, Yanjun Wang, Ziyu Chen, Linguo Li, Teng Li, Bingbing Ni, Wenjun Zhang
- abstract@open-review(Poster): 3D shape analysis has been widely explored in the era of deep learning. Numerous models have been developed for various 3D data representation formats, e.g., MeshCNN for meshes, PointNet for point clouds and VoxNet for voxels. In this study, we present Representation-Agnostic Shape Fields (RASF), a generalizable and computation-efficient shape embedding module for 3D deep learning. RASF is implemented with a learnable 3D grid with multiple channels to store local geometry. Based on RASF, shape embeddings for various 3D shape representations (point clouds, meshes and voxels) are retrieved by coordinate indexing. While there are multiple ways to optimize the learnable parameters of RASF, we provide two effective schemes among all in this paper for RASF pre-training: shape reconstruction and normal estimation. Once trained, RASF becomes a plug-and-play performance booster with negligible cost. Extensive experiments on diverse 3D representation formats, networks and applications, validate the universal effectiveness of the proposed RASF. Code and pre-trained models are publicly available\footnote{\url{https://github.com/seanywang0408/RASF}}.

Learning Synthetic Environments and Reward Networks for Reinforcement Learning

- Fabio Ferreira, Thomas Nierhoff, Andreas SÃ¤linger, Frank Hutter
- abstract@open-review(Poster): We introduce Synthetic Environments (SEs) and Reward Networks (RNs), represented by neural networks, as proxy environment models for training Reinforcement Learning (RL) agents. We show that an agent, after being trained exclusively on the SE, is able to solve the corresponding real environment. While an SE acts as a full proxy to a real environment by learning about its state dynamics and rewards, an RN is a partial proxy that learns to augment or replace rewards. We use bi-level optimization to evolve SEs and RNs: the inner loop trains the RL agent, and the outer loop trains the parameters of the SE / RN via an evolution strategy. We evaluate our proposed new concept on a broad range of RL algorithms and classic control environments. In a one-to-one comparison, learning an SE proxy requires more interactions with the real environment than training agents only on the real environment. However, once such an SE has been learned, we do not need any interactions with the real environment to train new agents. Moreover, the learned SE proxies allow us to train agents with fewer interactions while maintaining the original task performance. Our empirical results suggest that SEs achieve this result by learning informed representations that bias the agents towards relevant states. Moreover, we find that these proxies are robust against hyperparameter variation and can also transfer to unseen agents.

Who Is Your Right Mixup Partner in Positive and Unlabeled Learning

- Changchun Li, Ximing Li, Lei Feng, Jihong Ouyang
- abstract@open-review(Poster): Positive and Unlabeled (PU) learning targets inducing a binary classifier from weak training datasets of positive and unlabeled instances, which arise in many real-world applications. In this paper, we propose a novel PU learning method, namely Positive and unlabeled learning with Partially Positive Mixup (P3Mix), which simultaneously benefits from data augmentation and supervision correction with a heuristic mixup technique. To be specific, we take inspiration from the directional boundary deviation phenomenon observed in our preliminary experiments, where the learned PU boundary tends to deviate from the fully supervised boundary towards the positive side. For the unlabeled instances with ambiguous predictive results, we select their mixup partners from the positive instances around the learned PU boundary, so as to transform them into augmented instances near to the boundary yet with more precise supervision. Accordingly, those augmented instances may push the learned PU boundary towards the fully supervised boundary, thereby improving the classification performance. Comprehensive experimental results demonstrate the effectiveness of the heuristic mixup technique in PU learning and show that P3Mix can consistently outperform the state-of-the-art PU learning methods.

Incremental False Negative Detection for Contrastive Learning

- Tsai-Shien Chen, Wei-Chih Hung, Hung-Yu Tseng, Shao-Yi Chien, Ming-Hsuan Yang
- abstract@open-review(Poster): Self-supervised learning has recently shown great potential in vision tasks through contrastive learning, which aims to discriminate each image, or instance, in the dataset. However, such instance-level learning ignores the semantic relationship among instances and sometimes undesirably repels the anchor from the semantically similar samples, termed as "false negatives". In this work, we show that the unfavorable

effect from false negatives is more significant for the large-scale datasets with more semantic concepts. To address the issue, we propose a novel self-supervised contrastive learning framework that incrementally detects and explicitly removes the false negative samples. Specifically, following the training process, our method dynamically detects increasing high-quality false negatives considering that the encoder gradually improves and the embedding space becomes more semantically structural. Next, we discuss two strategies to explicitly remove the detected false negatives during contrastive learning. Extensive experiments show that our framework outperforms other self-supervised contrastive learning methods on multiple benchmarks in a limited resource setup.

Multi-Critic Actor Learning: Teaching RL Policies to Act with Style

- Siddharth Mysore, George Cheng, Yunqi Zhao, Kate Saenko, Meng Wu
- abstract@open-review(Poster): Using a single value function (critic) shared over multiple tasks in Actor-Critic multi-task reinforcement learning (MTRL) can result in negative interference between tasks, which can compromise learning performance. Multi-Critic Actor Learning (MultiCriticAL) proposes instead maintaining separate critics for each task being trained while training a single multi-task actor. Explicitly distinguishing between tasks also eliminates the need for critics to learn to do so and mitigates interference between task-value estimates. MultiCriticAL is tested in the context of multi-style learning, a special case of MTRL where agents are trained to behave with different distinct behavior styles, and yields up to 56% performance gains over the single-critic baselines and even successfully learns behavior styles in cases where single-critic approaches may simply fail to learn. In a simulated real-world use case, MultiCriticAL enables learning policies that smoothly transition between multiple fighting styles on an experimental build of EAâ€™s UFC game.

Clean Images are Hard to Reblur: Exploiting the Ill-Posed Inverse Task for Dynamic Scene Deblurring

- Seungjun Nah, Sanghyun Son, Jaerin Lee, Kyoung Mu Lee
- abstract@open-review(Poster): The goal of dynamic scene deblurring is to remove the motion blur in a given image. Typical learning-based approaches implement their solutions by minimizing the L1 or L2 distance between the output and the reference sharp image. Recent attempts adopt visual recognition features in training to improve the perceptual quality. However, those features are primarily designed to capture high-level contexts rather than low-level structures such as blurriness. Instead, we propose a more direct way to make images sharper by exploiting the inverse task of deblurring, namely, reblurring. Reblurring amplifies the remaining blur to rebuild the original blur, however, a well-deblurred clean image with zero-magnitude blur is hard to reblur. Thus, we design two types of reblurring loss functions for better deblurring. The supervised reblurring loss at training stage compares the amplified blur between the deblurred and the sharp images. The self-supervised reblurring loss at inference stage inspects if noticeable blur remains in the deblurred. Our experimental results on large-scale benchmarks and real images demonstrate the effectiveness of the reblurring losses in improving the perceptual quality of the deblurred images in terms of NIQE and LPIPS scores as well as visual sharpness.

Learning Disentangled Representation by Exploiting Pretrained Generative Models: A Contrastive Learning View

- Xuanchi Ren, Tao Yang, Yuwang Wang, Wenjun Zeng
- abstract@open-review(Poster): From the intuitive notion of disentanglement, the image variations corresponding to different generative factors should be distinct from each other, and the disentangled representation should reflect those variations with separate dimensions. To discover the generative factors and learn disentangled representation, previous methods typically leverage an extra regularization term when learning to generate realistic images. However, the term usually results in a trade-off between disentanglement and generation quality. For the generative models pretrained without any disentanglement term, the generated images show semantically meaningful variations when traversing along different directions in the latent space. Based on this observation, we argue that it is possible to mitigate the trade-off by (i) leveraging the pretrained generative models with high generation quality, (ii) focusing on discovering the traversal directions as generative factors for disentangled representation learning. To achieve this, we propose Disentaglement via Contrast (DisCo) as a framework to model the variations based on the target disentangled representations, and contrast the variations to jointly discover disentangled directions and learn disentangled representations. DisCo achieves the state-of-the-art disentangled representation learning and distinct direction discovering, given pretrained non-disentangled generative models including GAN, VAE, and Flow. Source code is at https://github.com/xrenaa/DisCo.

Towards Building A Group-based Unsupervised Representation Disentanglement Framework

- Tao Yang, Xuanchi Ren, Yuwang Wang, Wenjun Zeng, Nanning Zheng
- abstract@open-review(Poster): Disentangled representation learning is one of the major goals of deep learning, and is a key step for achieving explainable and generalizable models. The key idea of the state-of-the-art VAE-based unsupervised representation disentanglement methods is to minimize the total correlation of the joint distribution of the latent variables. However, it has been proved that their goal can not be achieved without introducing other inductive biases. The Group Theory based definition of representation disentanglement mathematically connects the data transformations to the representations using the formalism of group. In this paper, built on the group-based definition and inspired by the \emph{n-th dihedral group}, we first propose a theoretical framework towards achieving unsupervised representation disentanglement. We then propose a model based on existing VAE-based methods to tackle the unsupervised learning problem of the framework. In the theoretical framework, we prove three sufficient conditions on model, group structure, and data respectively in an effort to achieve, in an unsupervised way, disentangled representation per group-based definition. With these conditions, we offer an option, from the perspective of the group-based definition, for the inductive bias that existing VAE-based models lack. Experimentally, we train 1800 models covering the most prominent VAE-based methods on five datasets to verify the effectiveness of our theoretical framework. Compared to the original VAE-based methods, these Groupified VAEs consistently achieve better mean performance with smaller variances.

Learning Efficient Image Super-Resolution Networks via Structure-Regularized Pruning

- Yulun Zhang, Huan Wang, Can Qin, Yun Fu
- abstract@open-review(Poster): Several image super-resolution (SR) networks have been proposed of late for efficient SR, achieving promising results. However, they are still not lightweight enough and neglect to be extended to larger networks. At the same time, model compression techniques, like neural architecture search and knowledge distillation, typically consume considerable computation resources. In contrast, network pruning is a cheap and effective model compression technique. However, it is hard to be applied to SR networks directly because filter pruning for residual blocks is well-known tricky. To address the above issues, we propose structure-regularized pruning (SRP), which imposes regularization on the pruned structure to ensure the locations of pruned filters are aligned across different layers. Specifically, for the layers connected by the same residual, we select the filters of the same indices as unimportant filters. To transfer the expressive power in the unimportant filters to the rest of the network, we employ L_2 regularization to drive the weights towards zero so that eventually, their absence will cause minimal performance degradation. We apply SRP to train efficient image SR networks, resulting in a lightweight network SRPN-Lite and a very deep one SRPN. We conduct extensive comparisons with both lightweight and larger networks. SRPN-Lite and SRPN perform favorably against other recent efficient SR approaches quantitatively and visually.