Net-DNF: Effective Deep Modeling of Tabular Data

- Liran Katzir, Gal Elidan, Ran El-Yaniv
- abstract@open-review(Poster): A challenging open question in deep learning is how to handle tabular data. Unlike domains such as image and natural language processing, where deep architectures prevail, there is still no widely accepted neural architecture that dominates tabular data. As a step toward bridging this gap, we present Net-DNF a novel generic architecture whose inductive bias elicits models whose structure corresponds to logical Boolean formulas in disjunctive normal form (DNF) over affine soft-threshold decision terms. Net-DNFs also promote localized decisions that are taken over small subsets of the features. We present an extensive experiments showing that Net-DNFs significantly and consistently outperform fully connected networks over tabular data. With relatively few hyperparameters, Net-DNFs open the door to practical end-to-end handling of tabular data using neural networks. We present ablation studies, which justify the design choices of Net-DNF including the inductive bias elements, namely, Boolean formulation, locality, and feature selection.

Predicting Inductive Biases of Pre-Trained Models

- Charles Lovering, Rohan Jha, Tal Linzen, Ellie Pavlick
- abstract@open-review(Poster): Most current NLP systems are based on a pre-train-then-fine-tune paradigm, in which a large neural network is first trained in a self-supervised way designed to encourage the network to extract broadly-useful linguistic features, and then fine-tuned for a specific task of interest. Recent work attempts to understand why this recipe works and explain when it fails. Currently, such analyses have produced two sets of apparently-contradictory results. Work that analyzes the representations that result from pre-training (via "probing classifiers") finds evidence that rich features of linguistic structure can be decoded with high accuracy, but work that analyzes model behavior after fine-tuning (via "challenge sets") indicates that decisions are often not based on such structure but rather on spurious heuristics specific to the training set. In this work, we test the hypothesis that the extent to which a feature influences a model's decisions can be predicted using a combination of two factors: The feature's "extractability" after pre-training (measured using information-theoretic probing techniques), and the "evidence" available during fine-tuning (defined as the feature's co-occurrence rate with the label). In experiments with both synthetic and natural language data, we find strong evidence (statistically significant correlations) supporting this hypothesis.

Optimism in Reinforcement Learning with Generalized Linear Function Approximation

- Yining Wang, Ruosong Wang, Simon Shaolei Du, Akshay Krishnamurthy
- abstract@open-review(Poster): We design a new provably efficient algorithm for episodic reinforcement learning with generalized linear function approximation. We analyze the algorithm under a new expressivity assumption that we call ``optimistic closure," which is strictly weaker than assumptions from prior analyses for the linear setting. With optimistic closure, we prove that our algorithm enjoys a regret bound of $\widetilde{O}\left(H\sqrt{d^3 T}\right)$ where H is the horizon, d is the dimensionality of the state-action features and T is the number of episodes. This is the first statistically and computationally efficient algorithm for reinforcement learning with generalized linear functions.

SCoRe: Pre-Training for Context Representation in Conversational Semantic Parsing

- Tao Yu, Rui Zhang, Alex Polozov, Christopher Meek, Ahmed Hassan Awadallah
- abstract@open-review(Poster): Conversational Semantic Parsing (CSP) is the task of converting a sequence of natural language queries to formal language (e.g., SQL, SPARQL) that can be executed against a structured ontology (e.g. databases, knowledge bases). To accomplish this task, a CSP system needs to model the relation between the unstructured language utterance and the structured ontology while representing the multi-turn dynamics of the dialog. Pre-trained language models (LMs) are the state-of-the-art for various natural language processing tasks. However, existing pre-trained LMs that use language modeling training objectives over free-form text have limited ability to represent natural language references to contextual structural data. In this work, we present SCORE, a new pre-training approach for CSP tasks designed to induce representations that capture the alignment between the dialogue flow and the structural context. We demonstrate the broad applicability of SCORE to CSP tasks by combining SCORE with strong base systems on four different tasks (SPARC, COSQL, MWOZ, and SQA). We show that SCORE can improve the performance over all these base systems by a significant margin and achieves state-of-the-art results on three of them.

A teacher-student framework to distill future trajectories

- Alexander Neitz, Giambattista Parascandolo, Bernhard Schölkopf
- abstract@open-review(Poster): By learning to predict trajectories of dynamical systems, model-based methods can make extensive use of all observations from past experience. However, due to partial observability, stochasticity, compounding errors, and irrelevant dynamics, training to predict observations explicitly often results in poor models. Model-free techniques try to side-step the problem by learning to predict values directly. While breaking the explicit dependency on future observations can result in strong performance, this usually comes at the cost of low sample efficiency, as the abundant information about the dynamics contained in future observations goes unused. Here we take a step back from both approaches: Instead of hand-designing how trajectories should be incorporated, a teacher network learns to interpret the trajectories and to provide target activations which guide a student model that can only observe the present. The teacher is trained with meta-gradients to maximize the student's performance on a validation set. We show that our approach performs well on tasks that are difficult for model-free and model-based methods, and we study the role of every component through ablation studies.

Certify or Predict: Boosting Certified Robustness with Compositional Architectures

- Mark Niklas Mueller, Mislav Balunovic, Martin Vechev
- abstract@open-review(Poster): A core challenge with existing certified defense mechanisms is that while they improve certified robustness, they also tend to drastically decrease natural accuracy, making it difficult to use these methods in practice. In this work, we propose a new architecture which addresses this challenge and enables one to boost the certified robustness of any state-of-the-art deep network, while controlling the overall accuracy loss, without requiring retraining. The key idea is to combine this model with a (smaller) certified network where at inference time, an adaptive selection mechanism decides on the network to process the input sample. The approach is compositional: one can combine any pair of state-of-the-art (e.g., EfficientNet or ResNet) and certified networks, without restriction. The resulting architecture enables much higher natural accuracy than previously possible with certified defenses alone, while substantially boosting the certified robustness of deep networks. We demonstrate the effectiveness of this adaptive approach on a variety of datasets and architectures. For instance, on CIFAR-10 with an ℓ_∞ perturbation of 2/255, we are the first to obtain a high natural accuracy (90.1%) with non-trivial certified robustness (27.5%). Notably, prior state-of-the-art methods incur a substantial drop in accuracy for a similar certified robustness.

On the Transfer of Disentangled Representations in Realistic Settings

- Andrea Dittadi, Frederik Träuble, Francesco Locatello, Manuel Wuthrich, Vaibhav Agrawal, Ole Winther, Stefan Bauer, Bernhard Schölkopf
- abstract@open-review(Poster): Learning meaningful representations that disentangle the underlying structure of the data generating process is considered to be of key importance in machine learning. While disentangled representations were found to be useful for diverse tasks such as abstract reasoning and fair classification, their scalability and real-world impact remain questionable. We introduce a new high-resolution dataset with 1M simulated images and

over 1,800 annotated real-world images of the same setup. In contrast to previous work, this new dataset exhibits correlations, a complex underlying structure, and allows to evaluate transfer to unseen simulated and real-world settings where the encoder i) remains in distribution or ii) is out of distribution. We propose new architectures in order to scale disentangled representation learning to realistic high-resolution settings and conduct a large-scale empirical study of disentangled representations on this dataset. We observe that disentanglement is a good predictor for out-of-distribution (OOD) task performance.

Robust Reinforcement Learning on State Observations with Learned Optimal Adversary

- Huan Zhang, Hongge Chen, Duane S Boning, Cho-Jui Hsieh
- abstract@open-review(Poster): We study the robustness of reinforcement learning (RL) with adversarially perturbed state observations, which aligns with the setting of many adversarial attacks to deep reinforcement learning (DRL) and is also important for rolling out real-world RL agent under unpredictable sensing noise. With a fixed agent policy, we demonstrate that an optimal adversary to perturb state observations can be found, which is guaranteed to obtain the worst case agent reward. For DRL settings, this leads to a novel empirical adversarial attack to RL agents via a learned adversary that is much stronger than previous ones. To enhance the robustness of an agent, we propose a framework of alternating training with learned adversaries (ATLA), which trains an adversary online together with the agent using policy gradient following the optimal adversarial attack framework. Additionally, inspired by the analysis of state-adversarial Markov decision process (SA-MDP), we show that past states and actions (history) can be useful for learning a robust agent, and we empirically find a LSTM based policy can be more robust under adversaries. Empirical evaluations on a few continuous control environments show that ATLA achieves state-of-the-art performance under strong adversaries. Our code is available at https://github.com/huanzhang12/ATLA_robust_RL.

Practical Real Time Recurrent Learning with a Sparse Approximation

- Jacob Menick, Erich Elsen, Utku Evci, Simon Osindero, Karen Simonyan, Alex Graves
- abstract@open-review(Spotlight): Recurrent neural networks are usually trained with backpropagation through time, which requires storing a complete history of network states, and prohibits updating the weights "online" (after every timestep). Real Time Recurrent Learning (RTRL) eliminates the need for history storage and allows for online weight updates, but does so at the expense of computational costs that are quartic in the state size. This renders RTRL training intractable for all but the smallest networks, even ones that are made highly sparse. We introduce the Sparse n-step Approximation (SnAp) to the RTRL influence matrix. SnAp only tracks the influence of a parameter on hidden units that are reached by the computation graph within n timesteps of the recurrent core. SnAp with $n=1$ is no more expensive than backpropagation but allows training on arbitrarily long sequences. We find that it substantially outperforms other RTRL approximations with comparable costs such as Unbiased Online Recurrent Optimization. For highly sparse networks, SnAp with $n=2$ remains tractable and can outperform backpropagation through time in terms of learning speed when updates are done online.

Retrieval-Augmented Generation for Code Summarization via Hybrid GNN

- Shangqing Liu, Yu Chen, Xiaofei Xie, Jing Kai Siow, Yang Liu
- abstract@open-review(Spotlight): Source code summarization aims to generate natural language summaries from structured code snippets for better understanding code functionalities. However, automatic code summarization is challenging due to the complexity of the source code and the language gap between the source code and natural language summaries. Most previous approaches either rely on retrieval-based (which can take advantage of similar examples seen from the retrieval database, but have low generalization performance) or generation-based methods (which have better generalization performance, but cannot take advantage of similar examples). This paper proposes a novel retrieval-augmented mechanism to combine the benefits of both worlds. Furthermore, to mitigate the limitation of Graph Neural Networks (GNNs) on capturing global graph structure information of source code, we propose a novel attention-based dynamic graph to complement the static graph representation of the source code, and design a hybrid message passing GNN for capturing both the local and global structural information. To evaluate the proposed approach, we release a new challenging benchmark, crawled from diversified large-scale open-source C projects (total 95k+ unique functions in the dataset). Our method achieves the state-of-the-art performance, improving existing methods by 1.42, 2.44 and 1.29 in terms of BLEU-4, ROUGE-L and METEOR.

Learning from others' mistakes: Avoiding dataset biases without modeling them

- Victor Sanh, Thomas Wolf, Yonatan Belinkov, Alexander M Rush
- abstract@open-review(Poster): State-of-the-art natural language processing (NLP) models often learn to model dataset biases and surface form correlations instead of features that target the intended underlying task. Previous work has demonstrated effective methods to circumvent these issues when knowledge of the bias is available. We consider cases where the bias issues may not be explicitly identified, and show a method for training models that learn to ignore these problematic correlations. Our approach relies on the observation that models with limited capacity primarily learn to exploit biases in the dataset. We can leverage the errors of such limited capacity models to train a more robust model in a product of experts, thus bypassing the need to hand-craft a biased model. We show the effectiveness of this method to retain improvements in out-of-distribution settings even if no particular bias is targeted by the biased model.

Fast and Complete: Enabling Complete Neural Network Verification with Rapid and Massively Parallel Incomplete Verifiers

- Kaidi Xu, Huan Zhang, Shiqi Wang, Yihan Wang, Suman Jana, Xue Lin, Cho-Jui Hsieh
- abstract@open-review(Poster): Formal verification of neural networks (NNs) is a challenging and important problem. Existing efficient complete solvers typically require the branch-and-bound (BaB) process, which splits the problem domain into sub-domains and solves each sub-domain using faster but weaker incomplete verifiers, such as Linear Programming (LP) on linearly relaxed sub-domains. In this paper, we propose to use the backward mode linear relaxation based perturbation analysis (LiRPA) to replace LP during the BaB process, which can be efficiently implemented on the typical machine learning accelerators such as GPUs and TPUs. However, unlike LP, LiRPA when applied naively can produce much weaker bounds and even cannot check certain conflicts of sub-domains during splitting, making the entire procedure incomplete after BaB. To address these challenges, we apply a fast gradient based bound tightening procedure combined with batch splits and the design of minimal usage of LP bound procedure, enabling us to effectively use LiRPA on the accelerator hardware for the challenging complete NN verification problem and significantly outperform LP-based approaches. On a single GPU, we demonstrate an order of magnitude speedup compared to existing LP-based approaches.

Self-supervised Adversarial Robustness for the Low-label, High-data Regime

- Sven Gowal, Po-Sen Huang, Aaron van den Oord, Timothy Mann, Pushmeet Kohli
- abstract@open-review(Poster): Recent work discovered that training models to be invariant to adversarial perturbations requires substantially larger datasets than those required for standard classification. Perhaps more surprisingly, these larger datasets can be "mostly" unlabeled. Pseudo-labeling, a technique simultaneously pioneered by four separate and simultaneous works in 2019, has been proposed as a competitive alternative to labeled data for training adversarially robust models. However, when the amount of labeled data decreases, the performance of pseudo-labeling catastrophically drops, thus questioning the theoretical insights put forward by Uesato et al. (2019), which suggest that the sample complexity for learning an adversarially robust model from unlabeled data should match the fully supervised case. We introduce Bootstrap Your Own Robust Latents (BYORL), a self-supervised learning technique based on BYOL for training adversarially robust models. Our method enables us to train robust representations without any labels (reconciling practice with theory). Most notably, this robust representation can be leveraged by a linear classifier to train adversarially robust models, even when the linear classifier is not trained adversarially. We evaluate BYORL and pseudo-labeling on CIFAR-10 and ImageNet and demonstrate that BYORL

achieves significantly higher robustness (i.e., models resulting from BYORL are up to two times more accurate). Experiments on CIFAR-10 against ℓ_2 and ℓ_∞ norm-bounded perturbations demonstrate that BYORL achieves near state-of-the-art robustness with as little as 500 labeled examples. We also note that against ℓ_2 norm-bounded perturbations of size $\epsilon = 128/255$, BYORL surpasses the known state-of-the-art with an accuracy under attack of 77.61% (against 72.91% for the prior art).

Modeling the Second Player in Distributionally Robust Optimization

- Paul Michel, Tatsunori Hashimoto, Graham Neubig
- abstract@open-review(Poster): Distributionally robust optimization (DRO) provides a framework for training machine learning models that are able to perform well on a collection of related data distributions (the "uncertainty set"). This is done by solving a min-max game: the model is trained to minimize its maximum expected loss among all distributions in the uncertainty set. While careful design of the uncertainty set is critical to the success of the DRO procedure, previous work has been limited to relatively simple alternatives that keep the min-max optimization problem exactly tractable, such as f-divergence balls. In this paper, we argue instead for the use of neural generative models to characterize the worst-case distribution, allowing for more flexible and problem-specific selection of the uncertainty set. However, while simple conceptually, this approach poses a number of implementation and optimization challenges. To circumvent these issues, we propose a relaxation of the KL-constrained inner maximization objective that makes the DRO problem more amenable to gradient-based optimization of large scale generative models, and develop model selection heuristics to guide hyper-parameter search. On both toy settings and realistic NLP tasks, we find that the proposed approach yields models that are more robust than comparable baselines.

Neural Jump Ordinary Differential Equations: Consistent Continuous-Time Prediction and Filtering

- Calypso Herrera, Florian Krach, Josef Teichmann
- abstract@open-review(Poster): Combinations of neural ODEs with recurrent neural networks (RNN), like GRU-ODE-Bayes or ODE-RNN are well suited to model irregularly observed time series. While those models outperform existing discrete-time approaches, no theoretical guarantees for their predictive capabilities are available. Assuming that the irregularly-sampled time series data originates from a continuous stochastic process, the L^2-optimal online prediction is the conditional expectation given the currently available information. We introduce the Neural Jump ODE (NJ-ODE) that provides a data-driven approach to learn, continuously in time, the conditional expectation of a stochastic process. Our approach models the conditional expectation between two observations with a neural ODE and jumps whenever a new observation is made. We define a novel training framework, which allows us to prove theoretical guarantees for the first time. In particular, we show that the output of our model converges to the L^2-optimal prediction. This can be interpreted as solution to a special filtering problem. We provide experiments showing that the theoretical results also hold empirically. Moreover, we experimentally show that our model outperforms the baselines in more complex learning tasks and give comparisons on real-world datasets.

Gradient Origin Networks

- Sam Bond-Taylor, Chris G. Willcocks
- abstract@open-review(Poster): This paper proposes a new type of generative model that is able to quickly learn a latent representation without an encoder. This is achieved using empirical Bayes to calculate the expectation of the posterior, which is implemented by initialising a latent vector with zeros, then using the gradient of the log-likelihood of the data with respect to this zero vector as new latent points. The approach has similar characteristics to autoencoders, but with a simpler architecture, and is demonstrated in a variational autoencoder equivalent that permits sampling. This also allows implicit representation networks to learn a space of implicit functions without requiring a hypernetwork, retaining their representation advantages across datasets. The experiments show that the proposed method converges faster, with significantly lower reconstruction error than autoencoders, while requiring half the parameters.

On the mapping between Hopfield networks and Restricted Boltzmann Machines

- Matthew Smart, Anton Zilman
- abstract@open-review(Oral): Hopfield networks (HNs) and Restricted Boltzmann Machines (RBMs) are two important models at the interface of statistical physics, machine learning, and neuroscience. Recently, there has been interest in the relationship between HNs and RBMs, due to their similarity under the statistical mechanics formalism. An exact mapping between HNs and RBMs has been previously noted for the special case of orthogonal (“uncorrelated”) encoded patterns. We present here an exact mapping in the case of correlated pattern HNs, which are more broadly applicable to existing datasets. Specifically, we show that any HN with N binary variables and $p<N$ potentially correlated binary patterns can be transformed into an RBM with N binary visible variables and p gaussian hidden variables. We outline the conditions under which the reverse mapping exists, and conduct experiments on the MNIST dataset which suggest the mapping provides a useful initialization to the RBM weights. We discuss extensions, the potential importance of this correspondence for the training of RBMs, and for understanding the performance of feature extraction methods which utilize RBMs.

Efficient Generalized Spherical CNNs

- Oliver Cobb, Christopher G. R. Wallis, Augustine N. Mavor-Parker, Augustin Marignier, Matthew A. Price, Mayeul d'Avezac, Jason McEwen
- abstract@open-review(Poster): Many problems across computer vision and the natural sciences require the analysis of spherical data, for which representations may be learned efficiently by encoding equivariance to rotational symmetries. We present a generalized spherical CNN framework that encompasses various existing approaches and allows them to be leveraged alongside each other. The only existing non-linear spherical CNN layer that is strictly equivariant has complexity $\mathcal{O}(C^2L^5)$, where C is a measure of representational capacity and L the spherical harmonic bandlimit. Such a high computational cost often prohibits the use of strictly equivariant spherical CNNs. We develop two new strictly equivariant layers with reduced complexity $\mathcal{O}(CL^4)$ and $\mathcal{O}(CL^3 \log L)$, making larger, more expressive models computationally feasible. Moreover, we adopt efficient sampling theory to achieve further computational savings. We show that these developments allow the construction of more expressive hybrid models that achieve state-of-the-art accuracy and parameter efficiency on spherical benchmark problems.

DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION

- Pengcheng He, Xiaodong Liu, Jianfeng Gao, Weizhu Chen
- abstract@open-review(Poster): Recent progress in pre-trained neural language models has significantly improved the performance of many natural language processing (NLP) tasks. In this paper we propose a new model architecture DeBERTa (Decoding-enhanced BERT with disentangled attention) that improves the BERT and RoBERTa models using two novel techniques. The first is the disentangled attention mechanism, where each word is represented using two vectors that encode its content and position, respectively, and the attention weights among words are computed using disentangled matrices on their contents and relative positions, respectively. Second, an enhanced mask decoder is used to incorporate absolute positions in the decoding layer to predict the masked tokens in model pre-training. In addition, a new virtual adversarial training method is used for fine-tuning to improve models’ generalization. We show that these techniques significantly improve the efficiency of model pre-training and the performance of both natural language understand(NLU) and natural langauge generation (NLG) downstream tasks. Compared to RoBERTa-Large, a DeBERTa model trained on half of the training data performs consistently better on a wide range of NLP tasks, achieving improvements on MNLI by +0.9% (90.2% vs. 91.1%), on SQuAD v2.0 by +2.3% (88.4% vs. 90.7%) and RACE by +3.6% (83.2% vs. 86.8%). Notably, we scale up DeBERTa by training a larger version that consists of 48 Transform layers with 1.5 billion parameters. The significant performance boost makes the single DeBERTa model surpass the human performance on the SuperGLUE benchmark (Wang et al., 2019a) for the first time in terms of macro-average score (89.9 versus 89.8), and the ensemble DeBERTa model sits

atop the SuperGLUE leaderboard as of January 6, 2021, outperforming the human baseline by a decent margin (90.3 versus 89.8). The pre-trained DeBERTa models and the source code were released at: https://github.com/microsoft/DeBERTa.

Optimizing Memory Placement using Evolutionary Graph Reinforcement Learning

- Shauharda Khadka, Estelle Aflalo, Mattias Mardar, Avrech Ben-David, Santiago Miret, Shie Mannor, Tamir Hazan, Hanlin Tang, Somdeb Majumdar
- abstract@open-review(Poster): For deep neural network accelerators, memory movement is both energetically expensive and can bound computation. Therefore, optimal mapping of tensors to memory hierarchies is critical to performance. The growing complexity of neural networks calls for automated memory mapping instead of manual heuristic approaches; yet the search space of neural network computational graphs have previously been prohibitively large. We introduce Evolutionary Graph Reinforcement Learning (EGRL), a method designed for large search spaces, that combines graph neural networks, reinforcement learning, and evolutionary search. A set of fast, stateless policies guide the evolutionary search to improve its sample-efficiency. We train and validate our approach directly on the Intel NNP-I chip for inference. EGRL outperforms policy-gradient, evolutionary search and dynamic programming baselines on BERT, ResNet-101 and ResNet-50. We additionally achieve 28-78% speed-up compared to the native NNP-I compiler on all three workloads.

On the geometry of generalization and memorization in deep neural networks

- Cory Stephenson, suchismita padhy, Abhinav Ganesh, Yue Hui, Hanlin Tang, SueYeon Chung
- abstract@open-review(Poster): Understanding how large neural networks avoid memorizing training data is key to explaining their high generalization performance. To examine the structure of when and where memorization occurs in a deep network, we use a recently developed replica-based mean field theoretic geometric analysis method. We find that all layers preferentially learn from examples which share features, and link this behavior to generalization performance. Memorization predominately occurs in the deeper layers, due to decreasing object manifolds' radius and dimension, whereas early layers are minimally affected. This predicts that generalization can be restored by reverting the final few layer weights to earlier epochs before significant memorization occurred, which is confirmed by the experiments. Additionally, by studying generalization under different model sizes, we reveal the connection between the double descent phenomenon and the underlying model geometry. Finally, analytical analysis shows that networks avoid memorization early in training because close to initialization, the gradient contribution from permuted examples are small. These findings provide quantitative evidence for the structure of memorization across layers of a deep neural network, the drivers for such structure, and its connection to manifold geometric properties.

Continual learning in recurrent neural networks

- Benjamin Ehret, Christian Henning, Maria Cervera, Alexander Meulemans, Johannes Von Oswald, Benjamin F Grewe
- abstract@open-review(Poster): While a diverse collection of continual learning (CL) methods has been proposed to prevent catastrophic forgetting, a thorough investigation of their effectiveness for processing sequential data with recurrent neural networks (RNNs) is lacking. Here, we provide the first comprehensive evaluation of established CL methods on a variety of sequential data benchmarks. Specifically, we shed light on the particularities that arise when applying weight-importance methods, such as elastic weight consolidation, to RNNs. In contrast to feedforward networks, RNNs iteratively reuse a shared set of weights and require working memory to process input samples. We show that the performance of weight-importance methods is not directly affected by the length of the processed sequences, but rather by high working memory requirements, which lead to an increased need for stability at the cost of decreased plasticity for learning subsequent tasks. We additionally provide theoretical arguments supporting this interpretation by studying linear RNNs. Our study shows that established CL methods can be successfully ported to the recurrent case, and that a recent regularization approach based on hypernetworks outperforms weight-importance methods, thus emerging as a promising candidate for CL in RNNs. Overall, we provide insights on the differences between CL in feedforward networks and RNNs, while guiding towards effective solutions to tackle CL on sequential data.

Witches' Brew: Industrial Scale Data Poisoning via Gradient Matching

- Jonas Geiping, Liam H Fowl, W. Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, Tom Goldstein
- abstract@open-review(Poster): Data Poisoning attacks modify training data to maliciously control a model trained on such data. In this work, we focus on targeted poisoning attacks which cause a reclassification of an unmodified test image and as such breach model integrity. We consider a particularly malicious poisoning attack that is both `from scratch" and`clean label", meaning we analyze an attack that successfully works against new, randomly initialized models, and is nearly imperceptible to humans, all while perturbing only a small fraction of the training data. Previous poisoning attacks against deep neural networks in this setting have been limited in scope and success, working only in simplified settings or being prohibitively expensive for large datasets. The central mechanism of the new attack is matching the gradient direction of malicious examples. We analyze why this works, supplement with practical considerations. and show its threat to real-world practitioners, finding that it is the first poisoning method to cause targeted misclassification in modern deep networks trained from scratch on a full-sized, poisoned ImageNet dataset. Finally we demonstrate the limitations of existing defensive strategies against such an attack, concluding that data poisoning is a credible threat, even for large-scale deep learning systems.

Overfitting for Fun and Profit: Instance-Adaptive Data Compression

- Ties van Rozendaal, Iris AM Huijben, Taco Cohen
- abstract@open-review(Poster): Neural data compression has been shown to outperform classical methods in terms of RD performance, with results still improving rapidly. At a high level, neural compression is based on an autoencoder that tries to reconstruct the input instance from a (quantized) latent representation, coupled with a prior that is used to losslessly compress these latents. Due to limitations on model capacity and imperfect optimization and generalization, such models will suboptimally compress test data in general. However, one of the great strengths of learned compression is that if the test-time data distribution is known and relatively low-entropy (e.g. a camera watching a static scene, a dash cam in an autonomous car, etc.), the model can easily be finetuned or adapted to this distribution, leading to improved RD performance. In this paper we take this concept to the extreme, adapting the full model to a single video, and sending model updates (quantized and compressed using a parameter-space prior) along with the latent representation. Unlike previous work, we finetune not only the encoder/latents but the entire model, and - during finetuning - take into account both the effect of model quantization and the additional costs incurred by sending the model updates. We evaluate an image compression model on I-frames (sampled at 2 fps) from videos of the Xiph dataset, and demonstrate that full-model adaptation improves RD performance by ~1 dB, with respect to encoder-only finetuning.

A Block Minifloat Representation for Training Deep Neural Networks

- Sean Fox, Seyedramin Rasoulinezhad, Julian Faraone, david boland, Philip Leong
- abstract@open-review(Poster): Training Deep Neural Networks (DNN) with high efficiency can be difficult to achieve with native floating-point representations and commercially available hardware. Specialized arithmetic with custom acceleration offers perhaps the most promising alternative. Ongoing research is trending towards narrow floating-point representations, called minifloats, that pack more operations for a given silicon area and consume less power. In this paper, we introduce Block Minifloat (BM), a new spectrum of minifloat formats capable of training DNNs end-to-end with only 4-8 bit weight, activation and gradient tensors. While standard floating-point representations have two degrees of freedom, via the exponent and mantissa, BM exposes the exponent bias as an additional field for optimization. Crucially, this enables training with fewer exponent bits, yielding dense integer-like hardware for fused multiply-add (FMA) operations. For ResNet trained on ImageNet, 6-bit BM achieves almost no degradation in floating-

point accuracy with FMA units that are $4.1\times(23.9\times)$ smaller and consume $2.3\times(16.1\times)$ less energy than FP8 (FP32). Furthermore, our 8-bit BM format matches floating-point accuracy while delivering a higher computational density and faster expected training times.

Representation Learning via Invariant Causal Mechanisms

- Jovana Mitrovic, Brian McWilliams, Jacob C Walker, Lars Holger Buesing, Charles Blundell
- abstract@open-review(Poster): Self-supervised learning has emerged as a strategy to reduce the reliance on costly supervised signal by pretraining representations only using unlabeled data. These methods combine heuristic proxy classification tasks with data augmentations and have achieved significant success, but our theoretical understanding of this success remains limited. In this paper we analyze self-supervised representation learning using a causal framework. We show how data augmentations can be more effectively utilized through explicit invariance constraints on the proxy classifiers employed during pretraining. Based on this, we propose a novel self-supervised objective, Representation Learning via Invariant Causal Mechanisms (ReLIC), that enforces invariant prediction of proxy targets across augmentations through an invariance regularizer which yields improved generalization guarantees. Further, using causality we generalize contrastive learning, a particular kind of self-supervised method, and provide an alternative theoretical explanation for the success of these methods. Empirically, ReLIC significantly outperforms competing methods in terms of robustness and out-of-distribution generalization on ImageNet, while also significantly outperforming these methods on Atari achieving above human-level performance on 51 out of 57 games.

Sparse encoding for more-interpretable feature-selecting representations in probabilistic matrix factorization

- Joshua C Chang, Patrick Fletcher, Jungmin Han, Ted L Chang, Shashaank Vattikuti, Bart Desmet, Ayah Zirikly, Carson C Chow
- abstract@open-review(Poster): Dimensionality reduction methods for count data are critical to a wide range of applications in medical informatics and other fields where model interpretability is paramount. For such data, hierarchical Poisson matrix factorization (HPF) and other sparse probabilistic non-negative matrix factorization (NMF) methods are considered to be interpretable generative models. They consist of sparse transformations for decoding their learned representations into predictions. However, sparsity in representation decoding does not necessarily imply sparsity in the encoding of representations from the original data features. HPF is often incorrectly interpreted in the literature as if it possesses encoder sparsity. The distinction between decoder sparsity and encoder sparsity is subtle but important. Due to the lack of encoder sparsity, HPF does not possess the column-clustering property of classical NMF -- the factor loading matrix does not sufficiently define how each factor is formed from the original features. We address this deficiency by self-consistently enforcing encoder sparsity, using a generalized additive model (GAM), thereby allowing one to relate each representation coordinate to a subset of the original data features. In doing so, the method also gains the ability to perform feature selection. We demonstrate our method on simulated data and give an example of how encoder sparsity is of practical use in a concrete application of representing inpatient comorbidities in Medicare patients.

Mapping the Timescale Organization of Neural Language Models

- Hsiang-Yun Sherry Chien, Jinhan Zhang, Christopher Honey
- abstract@open-review(Poster): In the human brain, sequences of language input are processed within a distributed and hierarchical architecture, in which higher stages of processing encode contextual information over longer timescales. In contrast, in recurrent neural networks which perform natural language processing, we know little about how the multiple timescales of contextual information are functionally organized. Therefore, we applied tools developed in neuroscience to map the “processing timescales” of individual units within a word-level LSTM language model. This timescale-mapping method assigned long timescales to units previously found to track long-range syntactic dependencies. Additionally, the mapping revealed a small subset of the network (less than 15% of units) with long timescales and whose function had not previously been explored. We next probed the functional organization of the network by examining the relationship between the processing timescale of units and their network connectivity. We identified two classes of long-timescale units: “controller” units composed a densely interconnected subnetwork and strongly projected to the rest of the network, while “integrator” units showed the longest timescales in the network, and expressed projection profiles closer to the mean projection profile. Ablating integrator and controller units affected model performance at different positions within a sentence, suggesting distinctive functions of these two sets of units. Finally, we tested the generalization of these results to a character-level LSTM model and models with different architectures. In summary, we demonstrated a model-free technique for mapping the timescale organization in recurrent neural networks, and we applied this method to reveal the timescale and functional organization of neural language models

Neural networks with late-phase weights

- Johannes Von Oswald, Seijin Kobayashi, Joao Sacramento, Alexander Meulemans, Christian Henning, Benjamin F Grewe
- abstract@open-review(Poster): The largely successful method of training neural networks is to learn their weights using some variant of stochastic gradient descent (SGD). Here, we show that the solutions found by SGD can be further improved by ensembling a subset of the weights in late stages of learning. At the end of learning, we obtain back a single model by taking a spatial average in weight space. To avoid incurring increased computational costs, we investigate a family of low-dimensional late-phase weight models which interact multiplicatively with the remaining parameters. Our results show that augmenting standard models with late-phase weights improves generalization in established benchmarks such as CIFAR-10/100, ImageNet and enwik8. These findings are complemented with a theoretical analysis of a noisy quadratic problem which provides a simplified picture of the late phases of neural network learning.

Uncertainty-aware Active Learning for Optimal Bayesian Classifier

- Guang Zhao, Edward Dougherty, Byung-Jun Yoon, Francis Alexander, Xiaoning Qian
- abstract@open-review(Poster): For pool-based active learning, in each iteration a candidate training sample is chosen for labeling by optimizing an acquisition function. In Bayesian classification, expected Loss Reduction~(ELR) methods maximize the expected reduction in the classification error given a new labeled candidate based on a one-step-look-ahead strategy. ELR is the optimal strategy with a single query; however, since such myopic strategies cannot identify the long-term effect of a query on the classification error, ELR may get stuck before reaching the optimal classifier. In this paper, inspired by the mean objective cost of uncertainty (MOCU), a metric quantifying the uncertainty directly affecting the classification error, we propose an acquisition function based on a weighted form of MOCU. Similar to ELR, the proposed method focuses on the reduction of the uncertainty that pertains to the classification error. But unlike any other existing scheme, it provides the critical advantage that the resulting Bayesian active learning algorithm guarantees convergence to the optimal classifier of the true model. We demonstrate its performance with both synthetic and real-world datasets.

ResNet After All: Neural ODEs and Their Numerical Solution

- Katharina Ott, Prateek Katiyar, Philipp Hennig, Michael Tiemann
- abstract@open-review(Poster): A key appeal of the recently proposed Neural Ordinary Differential Equation (ODE) framework is that it seems to provide a continuous-time extension of discrete residual neural networks. As we show herein, though, trained Neural ODE models actually depend on the specific numerical method used during training. If the trained model is supposed to be a flow generated from an ODE, it should be possible to choose another numerical solver with equal or smaller numerical error without loss of performance. We observe that if training relies on a solver with overly coarse discretization, then testing with another solver of equal or smaller numerical error results in a sharp drop in accuracy. In such cases, the combination of vector field and numerical method cannot be interpreted as a flow generated from an ODE, which arguably poses a fatal breakdown of the Neural ODE concept. We observe, however, that there exists a critical step size beyond which the training yields a valid ODE vector field. We propose a method that

monitors the behavior of the ODE solver during training to adapt its step size, aiming to ensure a valid ODE without unnecessarily increasing computational cost. We verify this adaption algorithm on a common bench mark dataset as well as a synthetic dataset.

Benefit of deep learning with non-convex noisy gradient descent: Provable excess risk bound and superiority to kernel methods

- Taiji Suzuki, Shunta Akiyama
- abstract@open-review(Spotlight): Establishing a theoretical analysis that explains why deep learning can outperform shallow learning such as kernel methods is one of the biggest issues in the deep learning literature. Towards answering this question, we evaluate excess risk of a deep learning estimator trained by a noisy gradient descent with ridge regularization on a mildly overparameterized neural network, and discuss its superiority to a class of linear estimators that includes neural tangent kernel approach, random feature model, other kernel methods, k-NN estimator and so on. We consider a teacher-student regression model, and eventually show that {\it any} linear estimator can be outperformed by deep learning in a sense of the minimax optimal rate especially for a high dimension setting. The obtained excess bounds are so-called fast learning rate which is faster than $O(1/\sqrt{n})$ that is obtained by usual Rademacher complexity analysis. This discrepancy is induced by the non-convex geometry of the model and the noisy gradient descent used for neural network training provably reaches a near global optimal solution even though the loss landscape is highly non-convex. Although the noisy gradient descent does not employ any explicit or implicit sparsity inducing regularization, it shows a preferable generalization performance that dominates linear estimators.

Generalized Variational Continual Learning

- Noel Loo, Siddharth Swaroop, Richard E Turner
- abstract@open-review(Poster): Continual learning deals with training models on new tasks and datasets in an online fashion. One strand of research has used probabilistic regularization for continual learning, with two of the main approaches in this vein being Online Elastic Weight Consolidation (Online EWC) and Variational Continual Learning (VCL). VCL employs variational inference, which in other settings has been improved empirically by applying likelihood-tempering. We show that applying this modification to VCL recovers Online EWC as a limiting case, allowing for interpolation between the two approaches. We term the general algorithm Generalized VCL (GVCL). In order to mitigate the observed overpruning effect of VI, we take inspiration from a common multi-task architecture, neural networks with task-specific FiLM layers, and find that this addition leads to significant performance gains, specifically for variational methods. In the small-data regime, GVCL strongly outperforms existing baselines. In larger datasets, GVCL with FiLM layers outperforms or is competitive with existing baselines in terms of accuracy, whilst also providing significantly better calibration.

Mind the Gap when Conditioning Amortised Inference in Sequential Latent-Variable Models

- Justin Bayer, Maximilian Soelch, Atanas Mirchev, Baris Kayalibay, Patrick van der Smagt
- abstract@open-review(Poster): Amortised inference enables scalable learning of sequential latent-variable models (LVMs) with the evidence lower bound (ELBO). In this setting, variational posteriors are often only partially conditioned. While the true posteriors depend, e.g., on the entire sequence of observations, approximate posteriors are only informed by past observations. This mimics the Bayesian filter---a mixture of smoothing posteriors. Yet, we show that the ELBO objective forces partially-conditioned amortised posteriors to approximate products of smoothing posteriors instead. Consequently, the learned generative model is compromised. We demonstrate these theoretical findings in three scenarios: traffic flow, handwritten digits, and aerial vehicle dynamics. Using fully-conditioned approximate posteriors, performance improves in terms of generative modelling and multi-step prediction.

CausalWorld: A Robotic Manipulation Benchmark for Causal Structure and Transfer Learning

- Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Manuel Wuthrich, Yoshua Bengio, Bernhard Schölkopf, Stefan Bauer
- abstract@open-review(Poster): Despite recent successes of reinforcement learning (RL), it remains a challenge for agents to transfer learned skills to related environments. To facilitate research addressing this problem, we proposeCausalWorld, a benchmark for causal structure and transfer learning in a robotic manipulation environment. The environment is a simulation of an open-source robotic platform, hence offering the possibility of sim-to-real transfer. Tasks consist of constructing 3D shapes from a set of blocks - inspired by how children learn to build complex structures. The key strength of CausalWorld is that it provides a combinatorial family of such tasks with common causal structure and underlying factors (including, e.g., robot and object masses, colors, sizes). The user (or the agent) may intervene on all causal variables, which allows for fine-grained control over how similar different tasks (or task distributions) are. One can thus easily define training and evaluation distributions of a desired difficulty level, targeting a specific form of generalization (e.g., only changes in appearance or object mass). Further, this common parametrization facilitates defining curricula by interpolating between an initial and a target task. While users may define their own task distributions, we present eight meaningful distributions as concrete benchmarks, ranging from simple to very challenging, all of which require long-horizon planning as well as precise low-level motor control. Finally, we provide baseline results for a subset of these tasks on distinct training curricula and corresponding evaluation protocols, verifying the feasibility of the tasks in this benchmark.

Transformer protein language models are unsupervised structure learners

- Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, Alexander Rives
- abstract@open-review(Poster): Unsupervised contact prediction is central to uncovering physical, structural, and functional constraints for protein structure determination and design. For decades, the predominant approach has been to infer evolutionary constraints from a set of related sequences. In the past year, protein language models have emerged as a potential alternative, but performance has fallen short of state-of-the-art approaches in bioinformatics. In this paper we demonstrate that Transformer attention maps learn contacts from the unsupervised language modeling objective. We find the highest capacity models that have been trained to date already outperform a state-of-the-art unsupervised contact prediction pipeline, suggesting these pipelines can be replaced with a single forward pass of an end-to-end model.

Neural ODE Processes

- Alexander Norcliffe, Cristian Bodnar, Ben Day, Jacob Moss, Pietro Liò
- abstract@open-review(Poster): Neural Ordinary Differential Equations (NODEs) use a neural network to model the instantaneous rate of change in the state of a system. However, despite their apparent suitability for dynamics-governed time-series, NODEs present a few disadvantages. First, they are unable to adapt to incoming data-points, a fundamental requirement for real-time applications imposed by the natural direction of time. Second, time-series are often composed of a sparse set of measurements that could be explained by many possible underlying dynamics. NODEs do not capture this uncertainty. In contrast, Neural Processes (NPs) are a new class of stochastic processes providing uncertainty estimation and fast data-adaptation, but lack an explicit treatment of the flow of time. To address these problems, we introduce Neural ODE Processes (NDPs), a new class of stochastic processes determined by a distribution over Neural ODEs. By maintaining an adaptive data-dependent distribution over the underlying ODE, we show that our model can successfully capture the dynamics of low-dimensional systems from just a few data-points. At the same time, we demonstrate that NDPs scale up to challenging high-dimensional time-series with unknown latent dynamics such as rotating MNIST digits.

The role of Disentanglement in Generalisation

- Milton Llera Montero, Casimir JH Ludwig, Rui Ponte Costa, Gaurav Malhotra, Jeffrey Bowers

- abstract@open-review(Poster): Combinatorial generalisation — the ability to understand and produce novel combinations of familiar elements — is a core capacity of human intelligence that current AI systems struggle with. Recently, it has been suggested that learning disentangled representations may help address this problem. It is claimed that such representations should be able to capture the compositional structure of the world which can then be combined to support combinatorial generalisation. In this study, we systematically tested how the degree of disentanglement affects various forms of generalisation, including two forms of combinatorial generalisation that varied in difficulty. We trained three classes of variational autoencoders (VAEs) on two datasets on an unsupervised task by excluding combinations of generative factors during training. At test time we ask the models to reconstruct the missing combinations in order to measure generalisation performance. Irrespective of the degree of disentanglement, we found that the models supported only weak combinatorial generalisation. We obtained the same outcome when we directly input perfectly disentangled representations as the latents, and when we tested a model on a more complex task that explicitly required independent generative factors to be controlled. While learning disentangled representations does improve interpretability and sample efficiency in some downstream tasks, our results suggest that they are not sufficient for supporting more difficult forms of generalisation.

Learning to live with Dale's principle: ANNs with separate excitatory and inhibitory units

- Jonathan Cornford, Damjan Kalajdzievski, Marco Leite, Amélie Lamarquette, Dimitri Michael Kullmann, Blake Aaron Richards
- abstract@open-review(Poster): The units in artificial neural networks (ANNs) can be thought of as abstractions of biological neurons, and ANNs are increasingly used in neuroscience research. However, there are many important differences between ANN units and real neurons. One of the most notable is the absence of Dale's principle, which ensures that biological neurons are either exclusively excitatory or inhibitory. Dale's principle is typically left out of ANNs because its inclusion impairs learning. This is problematic, because one of the great advantages of ANNs for neuroscience research is their ability to learn complicated, realistic tasks. Here, by taking inspiration from feedforward inhibitory interneurons in the brain we show that we can develop ANNs with separate populations of excitatory and inhibitory units that learn just as well as standard ANNs. We call these networks Dale's ANNs (DANNs). We present two insights that enable DANNs to learn well: (1) DANNs are related to normalization schemes, and can be initialized such that the inhibition centres and standardizes the excitatory activity, (2) updates to inhibitory neuron parameters should be scaled using corrections based on the Fisher Information matrix. These results demonstrate how ANNs that respect Dale's principle can be built without sacrificing learning performance, which is important for future work using ANNs as models of the brain. The results may also have interesting implications for how inhibitory plasticity in the real brain operates.

SALD: Sign Agnostic Learning with Derivatives

- Matan Atzmon, Yaron Lipman
- abstract@open-review(Poster): Learning 3D geometry directly from raw data, such as point clouds, triangle soups, or unoriented meshes is still a challenging task that feeds many downstream computer vision and graphics applications.

In this paper, we introduce SALD: a method for learning implicit neural representations of shapes directly from raw data. We generalize sign agnostic learning (SAL) to include derivatives: given an unsigned distance function to the input raw data, we advocate a novel sign agnostic regression loss, incorporating both pointwise values and gradients of the unsigned distance function. Optimizing this loss leads to a signed implicit function solution, the zero level set of which is a high quality and valid manifold approximation to the input 3D data. The motivation behind SALD is that incorporating derivatives in a regression loss leads to a lower sample complexity, and consequently better fitting. In addition, we provide empirical evidence, as well as theoretical motivation in 2D that SAL enjoys a minimal surface property, favoring minimal area solutions. More importantly, we are able to show that this property still holds for SALD, i.e., with derivatives included.

We demonstrate the efficacy of SALD for shape space learning on two challenging datasets: ShapeNet that contains inconsistent orientation and non-manifold meshes, and D-Faust that contains raw 3D scans (triangle soups). On both these datasets, we present state-of-the-art results.

Ringing ReLUs: Harmonic Distortion Analysis of Nonlinear Feedforward Networks

- Christian H.X. Ali Mehmeti-Göpel, David Hartmann, Michael Wand
- abstract@open-review(Poster): In this paper, we apply harmonic distortion analysis to understand the effect of nonlinearities in the spectral domain. Each nonlinear layer creates higher-frequency harmonics, which we call "blueshift", whose magnitude increases with network depth, thereby increasing the “roughness” of the output landscape. Unlike differential models (such as vanishing gradients, sharpness), this provides a more global view of how network architectures behave across larger areas of their parameter domain. For example, the model predicts that residual connections are able to counter the effect by dampening corresponding higher frequency modes. We empirically verify the connection between blueshift and architectural choices, and provide evidence for a connection with trainability.

CoCon: A Self-Supervised Approach for Controlled Text Generation

- Alvin Chan, Yew-Soon Ong, Bill Pung, Aston Zhang, Jie Fu
- abstract@open-review(Poster): Pretrained Transformer-based language models (LMs) display remarkable natural language generation capabilities. With their immense potential, controlling text generation of such LMs is getting attention. While there are studies that seek to control high-level attributes (such as sentiment and topic) of generated text, there is still a lack of more precise control over its content at the word- and phrase-level. Here, we propose Content-Conditioner (CoCon) to control an LM's output text with a content input, at a fine-grained level. In our self-supervised approach, the CoCon block learns to help the LM complete a partially-observed text sequence by conditioning with content inputs that are withheld from the LM. Through experiments, we show that CoCon can naturally incorporate target content into generated texts and control high-level text attributes in a zero-shot manner.

Bidirectional Variational Inference for Non-Autoregressive Text-to-Speech

- Yoonhyung Lee, Joongbo Shin, Kyomin Jung
- abstract@open-review(Poster): Although early text-to-speech (TTS) models such as Tacotron 2 have succeeded in generating human-like speech, their autoregressive architectures have several limitations: (1) They require a lot of time to generate a mel-spectrogram consisting of hundreds of steps. (2) The autoregressive speech generation shows a lack of robustness due to its error propagation property. In this paper, we propose a novel non-autoregressive TTS model called BVAE-TTS, which eliminates the architectural limitations and generates a mel-spectrogram in parallel. BVAE-TTS adopts a bidirectional-inference variational autoencoder (BVAE) that learns hierarchical latent representations using both bottom-up and top-down paths to increase its expressiveness. To apply BVAE to TTS, we design our model to utilize text information via an attention mechanism. By using attention maps that BVAE-TTS generates, we train a duration predictor so that the model uses the predicted duration of each phoneme at inference. In experiments conducted on LJSpeech dataset, we show that our model generates a mel-spectrogram 27 times faster than Tacotron 2 with similar speech quality. Furthermore, our BVAE-TTS outperforms Glow-TTS, which is one of the state-of-the-art non-autoregressive TTS models, in terms of both speech quality and inference speed while having 58% fewer parameters.

Learning continuous-time PDEs from sparse data with graph neural networks

- Valerii Iakovlev, Markus Heinonen, Harri Lähdesmäki

- abstract@open-review(Poster): The behavior of many dynamical systems follow complex, yet still unknown partial differential equations (PDEs). While several machine learning methods have been proposed to learn PDEs directly from data, previous methods are limited to discrete-time approximations or make the limiting assumption of the observations arriving at regular grids. We propose a general continuous-time differential model for dynamical systems whose governing equations are parameterized by message passing graph neural networks. The model admits arbitrary space and time discretizations, which removes constraints on the locations of observation points and time intervals between the observations. The model is trained with continuous-time adjoint method enabling efficient neural PDE inference. We demonstrate the model's ability to work with unstructured grids, arbitrary time steps, and noisy observations. We compare our method with existing approaches on several well-known physical systems that involve first and higher-order PDEs with state-of-the-art predictive performance.

NAS-Bench-ASR: Reproducible Neural Architecture Search for Speech Recognition

- Abhinav Mehrotra, Alberto Gil C. P. Ramos, Sourav Bhattacharya, Łukasz Dudziak, Ravichander Vipperla, Thomas Chau, Mohamed S Abdelfattah, Samin Ishtiaq, Nicholas Donald Lane
- abstract@open-review(Poster): Powered by innovations in novel architecture design, noise tolerance techniques and increasing model capacity, Automatic Speech Recognition (ASR) has made giant strides in reducing word-error-rate over the past decade. ASR models are often trained with tens of thousand hours of high quality speech data to produce state-of-the-art (SOTA) results. Industry-scale ASR model training thus remains computationally heavy and time-consuming, and consequently has attracted little attention in adopting automatic techniques. On the other hand, Neural Architecture Search (NAS) has gained a lot of interest in the recent years thanks to its successes in discovering efficient architectures, often outperforming handcrafted alternatives. However, by changing the standard training process into a bi-level optimisation problem, NAS approaches often require significantly more time and computational power compared to single-model training, and at the same time increase complexity of the overall process. As a result, NAS has been predominately applied to problems which do not require as extensive training as ASR, and even then reproducibility of NAS algorithms is often problematic. Lately, a number of benchmark datasets has been introduced to address reproducibility issues by pro- viding NAS researchers with information about performance of different models obtained through exhaustive evaluation. However, these datasets focus mainly on computer vision and NLP tasks and thus suffer from limited coverage of application domains. In order to increase diversity in the existing NAS benchmarks, and at the same time provide systematic study of the effects of architectural choices for ASR, we release NAS-Bench-ASR – the first NAS benchmark for ASR models. The dataset consists of 8, 242 unique models trained on the TIMIT audio dataset for three different target epochs, and each starting from three different initializations. The dataset also includes runtime measurements of all the models on a diverse set of hardware platforms. Lastly, we show that identified good cell structures in our search space for TIMIT transfer well to a much larger LibriSpeech dataset.

Collective Robustness Certificates: Exploiting Interdependence in Graph Neural Networks

- Jan Schuchardt, Aleksandar Bojchevski, Johannes Gasteiger, Stephan Günnemann
- abstract@open-review(Poster): In tasks like node classification, image segmentation, and named-entity recognition we have a classifier that simultaneously outputs multiple predictions (a vector of labels) based on a single input, i.e. a single graph, image, or document respectively. Existing adversarial robustness certificates consider each prediction independently and are thus overly pessimistic for such tasks. They implicitly assume that an adversary can use different perturbed inputs to attack different predictions, ignoring the fact that we have a single shared input. We propose the first collective robustness certificate which computes the number of predictions that are simultaneously guaranteed to remain stable under perturbation, i.e. cannot be attacked. We focus on Graph Neural Networks and leverage their locality property - perturbations only affect the predictions in a close neighborhood - to fuse multiple single-node certificates into a drastically stronger collective certificate. For example, on the Citeseer dataset our collective certificate for node classification increases the average number of certifiable feature perturbations from 7 to 351.

Adversarially Guided Actor-Critic

- Yannis Flet-Berliac, Johan Ferret, Olivier Pietquin, Philippe Preux, Matthieu Geist
- abstract@open-review(Poster): Despite definite success in deep reinforcement learning problems, actor-critic algorithms are still confronted with sample inefficiency in complex environments, particularly in tasks where efficient exploration is a bottleneck. These methods consider a policy (the actor) and a value function (the critic) whose respective losses are built using different motivations and approaches. This paper introduces a third protagonist: the adversary. While the adversary mimics the actor by minimizing the KL-divergence between their respective action distributions, the actor, in addition to learning to solve the task, tries to differentiate itself from the adversary predictions. This novel objective stimulates the actor to follow strategies that could not have been correctly predicted from previous trajectories, making its behavior innovative in tasks where the reward is extremely rare. Our experimental analysis shows that the resulting Adversarially Guided Actor-Critic (AGAC) algorithm leads to more exhaustive exploration. Notably, AGAC outperforms current state-of-the-art methods on a set of various hard-exploration and procedurally-generated tasks.

Training independent subnetworks for robust prediction

- Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Zhe Liu, Jasper Snoek, Balaji Lakshminarayanan, Andrew Mingbo Dai, Dustin Tran
- abstract@open-review(Poster): Recent approaches to efficiently ensemble neural networks have shown that strong robustness and uncertainty performance can be achieved with a negligible gain in parameters over the original network. However, these methods still require multiple forward passes for prediction, leading to a significant runtime cost. In this work, we show a surprising result: the benefits of using multiple predictions can be achieved 'for free' under a single model's forward pass. In particular, we show that, using a multi-input multi-output (MIMO) configuration, one can utilize a single model's capacity to train multiple subnetworks that independently learn the task at hand. By ensembling the predictions made by the subnetworks, we improve model robustness without increasing compute. We observe a significant improvement in negative log-likelihood, accuracy, and calibration error on CIFAR10, CIFAR100, ImageNet, and their out-of-distribution variants compared to previous methods.

Complex Query Answering with Neural Link Predictors

- Erik Arakelyan, Daniel Daza, Pasquale Minervini, Michael Cochez
- abstract@open-review(Oral): Neural link predictors are immensely useful for identifying missing edges in large scale Knowledge Graphs. However, it is still not clear how to use these models for answering more complex queries that arise in a number of domains, such as queries using logical conjunctions ($\land$), disjunctions ($\lor$) and existential quantifiers ($\exists$), while accounting for missing edges. In this work, we propose a framework for efficiently answering complex queries on incomplete Knowledge Graphs. We translate each query into an end-to-end differentiable objective, where the truth value of each atom is computed by a pre-trained neural link predictor. We then analyse two solutions to the optimisation problem, including gradient-based and combinatorial search. In our experiments, the proposed approach produces more accurate results than state-of-the-art methods --- black-box neural models trained on millions of generated queries --- without the need of training on a large and diverse set of complex queries. Using orders of magnitude less training data, we obtain relative improvements ranging from 8% up to 40% in Hits@3 across different knowledge graphs containing factual information. Finally, we demonstrate that it is possible to explain the outcome of our model in terms of the intermediate solutions identified for each of the complex query atoms. All our source code and datasets are available online, at https://github.com/uclnlp/cqd.

Grounding Language to Autonomously-Acquired Skills via Goal Generation

- Ahmed Akakzia, Cédric Colas, Pierre-Yves Oudeyer, Mohamed CHETOUANI, Olivier Sigaud

- abstract@open-review(Poster): We are interested in the autonomous acquisition of repertoires of skills. Language-conditioned reinforcement learning (LC-RL) approaches are great tools in this quest, as they allow to express abstract goals as sets of constraints on the states. However, most LC-RL agents are not autonomous and cannot learn without external instructions and feedback. Besides, their direct language condition cannot account for the goal-directed behavior of pre-verbal infants and strongly limits the expression of behavioral diversity for a given language input. To resolve these issues, we propose a new conceptual approach to language-conditioned RL: the Language-Goal-Behavior architecture (LGB). LGB decouples skill learning and language grounding via an intermediate semantic representation of the world. To showcase the properties of LGB, we present a specific implementation called DECSTR. DECSTR is an intrinsically motivated learning agent endowed with an innate semantic representation describing spatial relations between physical objects. In a first stage G -> B, it freely explores its environment and targets self-generated semantic configurations. In a second stage (L -> G), it trains a language-conditioned goal generator to generate semantic goals that match the constraints expressed in language-based inputs. We showcase the additional properties of LGB w.r.t. both an end-to-end LC-RL approach and a similar approach leveraging non-semantic, continuous intermediate representations. Intermediate semantic representations help satisfy language commands in a diversity of ways, enable strategy switching after a failure and facilitate language grounding.

Hopfield Networks is All You Need

- Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K Kopp, Günter Klambauer, Johannes Brandstetter, Sepp Hochreiter
- abstract@open-review(Poster): We introduce a modern Hopfield network with continuous states and a corresponding update rule. The new Hopfield network can store exponentially (with the dimension of the associative space) many patterns, retrieves the pattern with one update, and has exponentially small retrieval errors. It has three types of energy minima (fixed points of the update): (1) global fixed point averaging over all patterns, (2) metastable states averaging over a subset of patterns, and (3) fixed points which store a single pattern. The new update rule is equivalent to the attention mechanism used in transformers. This equivalence enables a characterization of the heads of transformer models. These heads perform in the first layers preferably global averaging and in higher layers partial averaging via metastable states. The new modern Hopfield network can be integrated into deep learning architectures as layers to allow the storage of and access to raw input data, intermediate results, or learned prototypes. These Hopfield layers enable new ways of deep learning, beyond fully-connected, convolutional, or recurrent networks, and provide pooling, memory, association, and attention mechanisms. We demonstrate the broad applicability of the Hopfield layers across various domains. Hopfield layers improved state-of-the-art on three out of four considered multiple instance learning problems as well as on immune repertoire classification with several hundreds of thousands of instances. On the UCI benchmark collections of small classification tasks, where deep learning methods typically struggle, Hopfield layers yielded a new state-of-the-art when compared to different machine learning methods. Finally, Hopfield layers achieved state-of-the-art on two drug design datasets. The implementation is available at: \url{https://github.com/ml-jku/hopfield-layers}

Differentiable Trust Region Layers for Deep Reinforcement Learning

- Fabian Otto, Philipp Becker, Vien Anh Ngo, Hanna Carolin Maria Ziesche, Gerhard Neumann
- abstract@open-review(Poster): Trust region methods are a popular tool in reinforcement learning as they yield robust policy updates in continuous and discrete action spaces. However, enforcing such trust regions in deep reinforcement learning is difficult. Hence, many approaches, such as Trust Region Policy Optimization (TRPO) and Proximal Policy Optimization (PPO), are based on approximations. Due to those approximations, they violate the constraints or fail to find the optimal solution within the trust region. Moreover, they are difficult to implement, often lack sufficient exploration, and have been shown to depend on seemingly unrelated implementation choices. In this work, we propose differentiable neural network layers to enforce trust regions for deep Gaussian policies via closed-form projections. Unlike existing methods, those layers formalize trust regions for each state individually and can complement existing reinforcement learning algorithms. We derive trust region projections based on the Kullback-Leibler divergence, the Wasserstein L2 distance, and the Frobenius norm for Gaussian distributions. We empirically demonstrate that those projection layers achieve similar or better results than existing methods while being almost agnostic to specific implementation choices. The code is available at https://git.io/Jthb0.

Self-supervised Visual Reinforcement Learning with Object-centric Representations

- Andrii Zadaianchuk, Maximilian Seitzer, Georg Martius
- abstract@open-review(Spotlight): Autonomous agents need large repertoires of skills to act reasonably on new tasks that they have not seen before. However, acquiring these skills using only a stream of high-dimensional, unstructured, and unlabeled observations is a tricky challenge for any autonomous agent. Previous methods have used variational autoencoders to encode a scene into a low-dimensional vector that can be used as a goal for an agent to discover new skills. Nevertheless, in compositional/multi-object environments it is difficult to disentangle all the factors of variation into such a fixed-length representation of the whole scene. We propose to use object-centric representations as a modular and structured observation space, which is learned with a compositional generative world model. We show that the structure in the representations in combination with goal-conditioned attention policies helps the autonomous agent to discover and learn useful skills. These skills can be further combined to address compositional tasks like the manipulation of several different objects.

Temporally-Extended ε-Greedy Exploration

- Will Dabney, Georg Ostrovski, Andre Barreto
- abstract@open-review(Poster): Recent work on exploration in reinforcement learning (RL) has led to a series of increasingly complex solutions to the problem. This increase in complexity often comes at the expense of generality. Recent empirical studies suggest that, when applied to a broader set of domains, some sophisticated exploration methods are outperformed by simpler counterparts, such as ε-greedy. In this paper we propose an exploration algorithm that retains the simplicity of ε-greedy while reducing dithering. We build on a simple hypothesis: the main limitation of ε-greedy exploration is its lack of temporal persistence, which limits its ability to escape local optima. We propose a temporally extended form of ε-greedy that simply repeats the sampled action for a random duration. It turns out that, for many duration distributions, this suffices to improve exploration on a large set of domains. Interestingly, a class of distributions inspired by ecological models of animal foraging behaviour yields particularly strong performance.

Learning Associative Inference Using Fast Weight Memory

- Imanol Schlag, Tsendsuren Munkhdalai, Jürgen Schmidhuber
- abstract@open-review(Poster): Humans can quickly associate stimuli to solve problems in novel contexts. Our novel neural network model learns state representations of facts that can be composed to perform such associative inference. To this end, we augment the LSTM model with an associative memory, dubbed \textit{Fast Weight Memory} (FWM). Through differentiable operations at every step of a given input sequence, the LSTM \textit{updates and maintains} compositional associations stored in the rapidly changing FWM weights. Our model is trained end-to-end by gradient descent and yields excellent performance on compositional language reasoning problems, meta-reinforcement-learning for POMDPs, and small-scale word-level language modelling.

Multiscale Score Matching for Out-of-Distribution Detection

- Ahsan Mahmood, Junier Oliva, Martin Andreas Styner
- abstract@open-review(Poster): We present a new methodology for detecting out-of-distribution (OOD) images by utilizing norms of the score estimates at multiple noise scales. A score is defined to be the gradient of the log density with respect to the input data. Our methodology is completely unsupervised

and follows a straight forward training scheme. First, we train a deep network to estimate scores for L levels of noise. Once trained, we calculate the noisy score estimates for N in-distribution samples and take the L2-norms across the input dimensions (resulting in an NxL matrix). Then we train an auxiliary model (such as a Gaussian Mixture Model) to learn the in-distribution spatial regions in this L-dimensional space. This auxiliary model can now be used to identify points that reside outside the learned space. Despite its simplicity, our experiments show that this methodology significantly outperforms the state-of-the-art in detecting out-of-distribution images. For example, our method can effectively separate CIFAR-10 (inlier) and SVHN (OOD) images, a setting which has been previously shown to be difficult for deep likelihood models.

Learning to Sample with Local and Global Contexts in Experience Replay Buffer

- Youngmin Oh, Kimin Lee, Jinwoo Shin, Eunho Yang, Sung Ju Hwang
- abstract@open-review(Poster): Experience replay, which enables the agents to remember and reuse experience from the past, has played a significant role in the success of off-policy reinforcement learning (RL). To utilize the experience replay efficiently, the existing sampling methods allow selecting out more meaningful experiences by imposing priorities on them based on certain metrics (e.g. TD-error). However, they may result in sampling highly biased, redundant transitions since they compute the sampling rate for each transition independently, without consideration of its importance in relation to other transitions. In this paper, we aim to address the issue by proposing a new learning-based sampling method that can compute the relative importance of transition. To this end, we design a novel permutation-equivariant neural architecture that takes contexts from not only features of each transition (local) but also those of others (global) as inputs. We validate our framework, which we refer to as Neural Experience Replay Sampler (NERS), on multiple benchmark tasks for both continuous and discrete control tasks and show that it can significantly improve the performance of various off-policy RL methods. Further analysis confirms that the improvements of the sample efficiency indeed are due to sampling diverse and meaningful transitions by NERS that considers both local and global contexts.

Discovering a set of policies for the worst case reward

- Tom Zahavy, Andre Barreto, Daniel J Mankowitz, Shaobo Hou, Brendan O'Donoghue, Iurii Kemaev, Satinder Singh
- abstract@open-review(Spotlight): We study the problem of how to construct a set of policies that can be composed together to solve a collection of reinforcement learning tasks. Each task is a different reward function defined as a linear combination of known features. We consider a specific class of policy compositions which we call set improving policies (SIPs): given a set of policies and a set of tasks, a SIP is any composition of the former whose performance is at least as good as that of its constituents across all the tasks. We focus on the most conservative instantiation of SIPs, set-max policies (SMPs), so our analysis extends to any SIP. This includes known policy-composition operators like generalized policy improvement. Our main contribution is an algorithm that builds a set of policies in order to maximize the worst-case performance of the resulting SMP on the set of tasks. The algorithm works by successively adding new policies to the set. We show that the worst-case performance of the resulting SMP strictly improves at each iteration, and the algorithm only stops when there does not exist a policy that leads to improved performance. We empirically evaluate our algorithm on a grid world and also on a set of domains from the DeepMind control suite. We confirm our theoretical results regarding the monotonically improving performance of our algorithm. Interestingly, we also show empirically that the sets of policies computed by the algorithm are diverse, leading to different trajectories in the grid world and very distinct locomotion skills in the control suite.

Parameter-Based Value Functions

- Francesco Faccio, Louis Kirsch, Jürgen Schmidhuber
- abstract@open-review(Poster): Traditional off-policy actor-critic Reinforcement Learning (RL) algorithms learn value functions of a single target policy. However, when value functions are updated to track the learned policy, they forget potentially useful information about old policies. We introduce a class of value functions called Parameter-Based Value Functions (PBVFs) whose inputs include the policy parameters. They can generalize across different policies. PBVFs can evaluate the performance of any policy given a state, a state-action pair, or a distribution over the RL agent's initial states. First we show how PBVFs yield novel off-policy policy gradient theorems. Then we derive off-policy actor-critic algorithms based on PBVFs trained by Monte Carlo or Temporal Difference methods. We show how learned PBVFs can zero-shot learn new policies that outperform any policy seen during training. Finally our algorithms are evaluated on a selection of discrete and continuous control tasks using shallow policies and deep neural networks. Their performance is comparable to state-of-the-art methods.

New Bounds For Distributed Mean Estimation and Variance Reduction

- Peter Davies, Vijaykrishna Gurunanthan, Niusha Moshrefi, Saleh Ashkboos, Dan Alistarh
- abstract@open-review(Poster): We consider the problem of distributed mean estimation (DME), in which n machines are each given a local d-dimensional vector $\mathbf x_v \in \mathbb R^d$, and must cooperate to estimate the mean of their inputs $\mathbf \mu = \frac 1n\sum_{v = 1}^n \mathbf x_v$, while minimizing total communication cost. DME is a fundamental construct in distributed machine learning, and there has been considerable work on variants of this problem, especially in the context of distributed variance reduction for stochastic gradients in parallel SGD. Previous work typically assumes an upper bound on the norm of the input vectors, and achieves an error bound in terms of this norm. However, in many real applications, the input vectors are concentrated around the correct output $\mathbf \mu$, but $\mathbf \mu$ itself has large norm. In such cases, previous output error bounds perform poorly. In this paper, we show that output error bounds need not depend on input norm. We provide a method of quantization which allows distributed mean estimation to be performed with solution quality dependent only on the distance between inputs, not on input norm, and show an analogous result for distributed variance reduction. The technique is based on a new connection with lattice theory. We also provide lower bounds showing that the communication to error trade-off of our algorithms is asymptotically optimal. As the lattices achieving optimal bounds under ℓ_2-norm can be computationally impractical, we also present an extension which leverages easy-to-use cubic lattices, and is loose only up to a logarithmic factor in d. We show experimentally that our method yields practical improvements for common applications, relative to prior approaches.

Learning to Set Waypoints for Audio-Visual Navigation

- Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, Kristen Grauman
- abstract@open-review(Poster): In audio-visual navigation, an agent intelligently travels through a complex, unmapped 3D environment using both sights and sounds to find a sound source (e.g., a phone ringing in another room). Existing models learn to act at a fixed granularity of agent motion and rely on simple recurrent aggregations of the audio observations. We introduce a reinforcement learning approach to audio-visual navigation with two key novel elements: 1) waypoints that are dynamically set and learned end-to-end within the navigation policy, and 2) an acoustic memory that provides a structured, spatially grounded record of what the agent has heard as it moves. Both new ideas capitalize on the synergy of audio and visual data for revealing the geometry of an unmapped space. We demonstrate our approach on two challenging datasets of real-world 3D scenes, Replica and Matterport3D. Our model improves the state of the art by a substantial margin, and our experiments reveal that learning the links between sights, sounds, and space is essential for audio-visual navigation.

Disambiguating Symbolic Expressions in Informal Documents

- Dennis Müller, Cezary Kaliszyk
- abstract@open-review(Poster): We propose the task of \emph{disambiguating} symbolic expressions in informal STEM documents in the form of \LaTeX files -- that is, determining their precise semantics and abstract syntax tree -- as a neural machine translation task. We discuss the distinct challenges involved and present a dataset with roughly 33,000 entries. We evaluated several baseline models on this dataset, which failed to yield even syntactically

valid \LaTeX before overfitting. Consequently, we describe a methodology using a \emph{transformer} language model pre-trained on sources obtained from \url{arxiv.org}, which yields promising results despite the small size of the dataset. We evaluate our model using a plurality of dedicated techniques, taking syntax and semantics of symbolic expressions into account.

Colorization Transformer

- Manoj Kumar, Dirk Weissenborn, Nal Kalchbrenner
- abstract@open-review(Poster): We present the Colorization Transformer, a novel approach for diverse high fidelity image colorization based on self-attention. Given a grayscale image, the colorization proceeds in three steps. We first use a conditional autoregressive transformer to produce a low resolution coarse coloring of the grayscale image. Our architecture adopts conditional transformer layers to effectively condition grayscale input. Two subsequent fully parallel networks upsample the coarse colored low resolution image into a finely colored high resolution image. Sampling from the Colorization Transformer produces diverse colorings whose fidelity outperforms the previous state-of-the-art on colorising ImageNet based on FID results and based on a human evaluation in a Mechanical Turk test. Remarkably, in more than 60\% of cases human evaluators prefer the highest rated among three generated colorings over the ground truth. The code and pre-trained checkpoints for Colorization Transformer are publicly available at https://github.com/google-research/google-research/tree/master/coltran

Theoretical bounds on estimation error for meta-learning

- James Lucas, Mengye Ren, Irene Raissa KAMENI KAMENI, Toniann Pitassi, Richard Zemel
- abstract@open-review(Poster): Machine learning models have traditionally been developed under the assumption that the training and test distributions match exactly. However, recent success in few-shot learning and related problems are encouraging signs that these models can be adapted to more realistic settings where train and test distributions differ. Unfortunately, there is severely limited theoretical support for these algorithms and little is known about the difficulty of these problems. In this work, we provide novel information-theoretic lower-bounds on minimax rates of convergence for algorithms that are trained on data from multiple sources and tested on novel data. Our bounds depend intuitively on the information shared between sources of data, and characterize the difficulty of learning in this setting for arbitrary algorithms. We demonstrate these bounds on a hierarchical Bayesian model of meta-learning, computing both upper and lower bounds on parameter estimation via maximum-a-posteriori inference.

Implicit Normalizing Flows

- Cheng Lu, Jianfei Chen, Chongxuan Li, Qiuhao Wang, Jun Zhu
- abstract@open-review(Spotlight): Normalizing flows define a probability distribution by an explicit invertible transformation $\boldsymbol{\mathbf{z}}=f(\boldsymbol{\mathbf{x}})$. In this work, we present implicit normalizing flows (ImpFlows), which generalize normalizing flows by allowing the mapping to be implicitly defined by the roots of an equation $F(\boldsymbol{\mathbf{z}}, \boldsymbol{\mathbf{x}})= \boldsymbol{\mathbf{0}}$. ImpFlows build on residual flows (ResFlows) with a proper balance between expressiveness and tractability. Through theoretical analysis, we show that the function space of ImpFlow is strictly richer than that of ResFlows. Furthermore, for any ResFlow with a fixed number of blocks, there exists some function that ResFlow has a non-negligible approximation error. However, the function is exactly representable by a single-block ImpFlow. We propose a scalable algorithm to train and draw samples from ImpFlows. Empirically, we evaluate ImpFlow on several classification and density modeling tasks, and ImpFlow outperforms ResFlow with a comparable amount of parameters on all the benchmarks.

Variational Information Bottleneck for Effective Low-Resource Fine-Tuning

- Rabeeh Karimi mahabadi, Yonatan Belinkov, James Henderson
- abstract@open-review(Poster): While large-scale pretrained language models have obtained impressive results when fine-tuned on a wide variety of tasks, they still often suffer from overfitting in low-resource scenarios. Since such models are general-purpose feature extractors, many of these features are inevitably irrelevant for a given target task. We propose to use Variational Information Bottleneck (VIB) to suppress irrelevant features when fine-tuning on low-resource target tasks, and show that our method successfully reduces overfitting. Moreover, we show that our VIB model finds sentence representations that are more robust to biases in natural language inference datasets, and thereby obtains better generalization to out-of-domain datasets. Evaluation on seven low-resource datasets in different tasks shows that our method significantly improves transfer learning in low-resource scenarios, surpassing prior work. Moreover, it improves generalization on 13 out of 15 out-of-domain natural language inference benchmarks. Our code is publicly available in https://github.com/rabeehk/vibert.

TropEx: An Algorithm for Extracting Linear Terms in Deep Neural Networks

- Martin Trimmel, Henning Petzka, Cristian Sminchisescu
- abstract@open-review(Poster): Deep neural networks with rectified linear (ReLU) activations are piecewise linear functions, where hyperplanes partition the input space into an astronomically high number of linear regions. Previous work focused on counting linear regions to measure the network's expressive power and on analyzing geometric properties of the hyperplane configurations. In contrast, we aim to understand the impact of the linear terms on network performance, by examining the information encoded in their coefficients. To this end, we derive TropEx, a nontrivial tropical algebra-inspired algorithm to systematically extract linear terms based on data. Applied to convolutional and fully-connected networks, our algorithm uncovers significant differences in how the different networks utilize linear regions for generalization. This underlines the importance of systematic linear term exploration, to better understand generalization in neural networks trained with complex data sets.

Seq2Tens: An Efficient Representation of Sequences by Low-Rank Tensor Projections

- Csaba Toth, Patric Bonnier, Harald Oberhauser
- abstract@open-review(Poster): Sequential data such as time series, video, or text can be challenging to analyse as the ordered structure gives rise to complex dependencies. At the heart of this is non-commutativity, in the sense that reordering the elements of a sequence can completely change its meaning. We use a classical mathematical object -- the free algebra -- to capture this non-commutativity. To address the innate computational complexity of this algebra, we use compositions of low-rank tensor projections. This yields modular and scalable building blocks that give state-of-the-art performance on standard benchmarks such as multivariate time series classification, mortality prediction and generative models for video.

Representation learning for improved interpretability and classification accuracy of clinical factors from EEG

- Garrett Honke, Irina Higgins, Nina Thigpen, Vladimir Miskovic, Katie Link, Sunny Duan, Pramod Gupta, Julia Klawohn, Greg Hajcak
- abstract@open-review(Poster): Despite extensive standardization, diagnostic interviews for mental health disorders encompass substantial subjective judgment. Previous studies have demonstrated that EEG-based neural measures can function as reliable objective correlates of depression, or even predictors of depression and its course. However, their clinical utility has not been fully realized because of 1) the lack of automated ways to deal with the inherent noise associated with EEG data at scale, and 2) the lack of knowledge of which aspects of the EEG signal may be markers of a clinical disorder. Here we adapt an unsupervised pipeline from the recent deep representation learning literature to address these problems by 1) learning a disentangled representation using β-VAE to denoise the signal, and 2) extracting interpretable features associated with a sparse set of clinical labels using a Symbol-Concept Association Network (SCAN). We demonstrate that our method is able to outperform the canonical hand-engineered baseline

classification method on a number of factors, including participant age and depression diagnosis. Furthermore, our method recovers a representation that can be used to automatically extract denoised Event Related Potentials (ERPs) from novel, single EEG trajectories, and supports fast supervised re-mapping to various clinical labels, allowing clinicians to re-use a single EEG representation regardless of updates to the standardized diagnostic system. Finally, single factors of the learned disentangled representations often correspond to meaningful markers of clinical factors, as automatically detected by SCAN, allowing for human interpretability and post-hoc expert analysis of the recommendations made by the model.

Language-Agnostic Representation Learning of Source Code from Structure and Context

- Daniel Zügner, Tobias Kirschstein, Michele Catasta, Jure Leskovec, Stephan Günnemann
- abstract@open-review(Poster): Source code (Context) and its parsed abstract syntax tree (AST; Structure) are two complementary representations of the same computer program. Traditionally, designers of machine learning models have relied predominantly either on Structure or Context. We propose a new model, which jointly learns on Context and Structure of source code. In contrast to previous approaches, our model uses only language-agnostic features, i.e., source code and features that can be computed directly from the AST. Besides obtaining state-of-the-art on monolingual code summarization on all five programming languages considered in this work, we propose the first multilingual code summarization model. We show that jointly training on non-parallel data from multiple programming languages improves results on all individual languages, where the strongest gains are on low-resource languages. Remarkably, multilingual training only from Context does not lead to the same improvements, highlighting the benefits of combining Structure and Context for representation learning on code.

Generalized Multimodal ELBO

- Thomas M. Sutter, Imant Daunhawer, Julia E Vogt
- abstract@open-review(Poster): Multiple data types naturally co-occur when describing real-world phenomena and learning from them is a long-standing goal in machine learning research. However, existing self-supervised generative models approximating an ELBO are not able to fulfill all desired requirements of multimodal models: their posterior approximation functions lead to a trade-off between the semantic coherence and the ability to learn the joint data distribution. We propose a new, generalized ELBO formulation for multimodal data that overcomes these limitations. The new objective encompasses two previous methods as special cases and combines their benefits without compromises. In extensive experiments, we demonstrate the advantage of the proposed method compared to state-of-the-art models in self-supervised, generative learning tasks.

Model-based micro-data reinforcement learning: what are the crucial model properties and which model to choose?

- Balázs Kégl, Gabriel Hurtado, Albert Thomas
- abstract@open-review(Poster): We contribute to micro-data model-based reinforcement learning (MBRL) by rigorously comparing popular generative models using a fixed (random shooting) control agent. We find that on an environment that requires multimodal posterior predictives, mixture density nets outperform all other models by a large margin. When multimodality is not required, our surprising finding is that we do not need probabilistic posterior predictives: deterministic models are on par, in fact they consistently (although non-significantly) outperform their probabilistic counterparts. We also found that heteroscedasticity at training time, perhaps acting as a regularizer, improves predictions at longer horizons. At the methodological side, we design metrics and an experimental protocol which can be used to evaluate the various models, predicting their asymptotic performance when using them on the control problem. Using this framework, we improve the state-of-the-art sample complexity of MBRL on Acrobot by two to four folds, using an aggressive training schedule which is outside of the hyperparameter interval usually considered.

Set Prediction without Imposing Structure as Conditional Density Estimation

- David W Zhang, Gertjan J. Burghouts, Cees G. M. Snoek
- abstract@open-review(Poster): Set prediction is about learning to predict a collection of unordered variables with unknown interrelations. Training such models with set losses imposes the structure of a metric space over sets. We focus on stochastic and underdefined cases, where an incorrectly chosen loss function leads to implausible predictions. Example tasks include conditional point-cloud reconstruction and predicting future states of molecules. In this paper we propose an alternative to training via set losses, by viewing learning as conditional density estimation. Our learning framework fits deep energy-based models and approximates the intractable likelihood with gradient-guided sampling. Furthermore, we propose a stochastically augmented prediction algorithm that enables multiple predictions, reflecting the possible variations in the target set. We empirically demonstrate on a variety of datasets the capability to learn multi-modal densities and produce different plausible predictions. Our approach is competitive with previous set prediction models on standard benchmarks. More importantly, it extends the family of addressable tasks beyond those that have unambiguous predictions.

Learning Value Functions in Deep Policy Gradients using Residual Variance

- Yannis Flet-Berliac, reda ouhamma, odalric-ambrym maillard, Philippe Preux
- abstract@open-review(Poster): Policy gradient algorithms have proven to be successful in diverse decision making and control tasks. However, these methods suffer from high sample complexity and instability issues. In this paper, we address these challenges by providing a different approach for training the critic in the actor-critic framework. Our work builds on recent studies indicating that traditional actor-critic algorithms do not succeed in fitting the true value function, calling for the need to identify a better objective for the critic. In our method, the critic uses a new state-value (resp. state-action-value) function approximation that learns the value of the states (resp. state-action pairs) relative to their mean value rather than the absolute value as in conventional actor-critic. We prove the theoretical consistency of the new gradient estimator and observe dramatic empirical improvement across a variety of continuous control tasks and algorithms. Furthermore, we validate our method in tasks with sparse rewards, where we provide experimental evidence and theoretical insights.

IDF++: Analyzing and Improving Integer Discrete Flows for Lossless Compression

- Rianne van den Berg, Alexey A. Gritsenko, Mostafa Dehghani, Casper Kaae Sønderby, Tim Salimans
- abstract@open-review(Poster): In this paper we analyse and improve integer discrete flows for lossless compression. Integer discrete flows are a recently proposed class of models that learn invertible transformations for integer-valued random variables. Their discrete nature makes them particularly suitable for lossless compression with entropy coding schemes. We start by investigating a recent theoretical claim that states that invertible flows for discrete random variables are less flexible than their continuous counterparts. We demonstrate with a proof that this claim does not hold for integer discrete flows due to the embedding of data with finite support into the countably infinite integer lattice. Furthermore, we zoom in on the effect of gradient bias due to the straight-through estimator in integer discrete flows, and demonstrate that its influence is highly dependent on architecture choices and less prominent than previously thought. Finally, we show how different architecture modifications improve the performance of this model class for lossless compression, and that they also enable more efficient compression: a model with half the number of flow layers performs on par with or better than the original integer discrete flow model.

Fully Unsupervised Diversity Denoising with Convolutional Variational Autoencoders

- Mangal Prakash, Alexander Krull, Florian Jug

- abstract@open-review(Poster): Deep Learning based methods have emerged as the indisputable leaders for virtually all image restoration tasks. Especially in the domain of microscopy images, various content-aware image restoration (CARE) approaches are now used to improve the interpretability of acquired data. Naturally, there are limitations to what can be restored in corrupted images, and like for all inverse problems, many potential solutions exist, and one of them must be chosen. Here, we propose DivNoising, a denoising approach based on fully convolutional variational autoencoders (VAEs), overcoming the problem of having to choose a single solution by predicting a whole distribution of denoised images. First we introduce a principled way of formulating the unsupervised denoising problem within the VAE framework by explicitly incorporating imaging noise models into the decoder. Our approach is fully unsupervised, only requiring noisy images and a suitable description of the imaging noise distribution. We show that such a noise model can either be measured, bootstrapped from noisy data, or co-learned during training. If desired, consensus predictions can be inferred from a set of DivNoising predictions, leading to competitive results with other unsupervised methods and, on occasion, even with the supervised state-of-the-art. DivNoising samples from the posterior enable a plethora of useful applications. We are (i) showing denoising results for 13 datasets, (ii) discussing how optical character recognition (OCR) applications can benefit from diverse predictions, and are (iii) demonstrating how instance cell segmentation improves when using diverse DivNoising predictions.

Is Attention Better Than Matrix Decomposition?

- Zhengyang Geng, Meng-Hao Guo, Hongxu Chen, Xia Li, Ke Wei, Zhouchen Lin
- abstract@open-review(Poster): As an essential ingredient of modern deep learning, attention mechanism, especially self-attention, plays a vital role in the global correlation discovery. However, is hand-crafted attention irreplaceable when modeling the global context? Our intriguing finding is that self-attention is not better than the matrix decomposition~(MD) model developed 20 years ago regarding the performance and computational cost for encoding the long-distance dependencies. We model the global context issue as a low-rank completion problem and show that its optimization algorithms can help design global information blocks. This paper then proposes a series of Hamburgers, in which we employ the optimization algorithms for solving MDs to factorize the input representations into sub-matrices and reconstruct a low-rank embedding. Hamburgers with different MDs can perform favorably against the popular global context module self-attention when carefully coping with gradients back-propagated through MDs. Comprehensive experiments are conducted in the vision tasks where it is crucial to learn the global context, including semantic segmentation and image generation, demonstrating significant improvements over self-attention and its variants. Code is available at https://github.com/Gsunshine/Enjoy-Hamburger.

Improving Transformation Invariance in Contrastive Representation Learning

- Adam Foster, Rattana Pukdee, Tom Rainforth
- abstract@open-review(Poster): We propose methods to strengthen the invariance properties of representations obtained by contrastive learning. While existing approaches implicitly induce a degree of invariance as representations are learned, we look to more directly enforce invariance in the encoding process. To this end, we first introduce a training objective for contrastive learning that uses a novel regularizer to control how the representation changes under transformation. We show that representations trained with this objective perform better on downstream tasks and are more robust to the introduction of nuisance transformations at test time. Second, we propose a change to how test time representations are generated by introducing a feature averaging approach that combines encodings from multiple transformations of the original input, finding that this leads to across the board performance gains. Finally, we introduce the novel Spirograph dataset to explore our ideas in the context of a differentiable generative process with multiple downstream tasks, showing that our techniques for learning invariance are highly beneficial.

On the Origin of Implicit Regularization in Stochastic Gradient Descent

- Samuel L Smith, Benoit Dherin, David Barrett, Soham De
- abstract@open-review(Poster): For infinitesimal learning rates, stochastic gradient descent (SGD) follows the path of gradient flow on the full batch loss function. However moderately large learning rates can achieve higher test accuracies, and this generalization benefit is not explained by convergence bounds, since the learning rate which maximizes test accuracy is often larger than the learning rate which minimizes training loss. To interpret this phenomenon we prove that for SGD with random shuffling, the mean SGD iterate also stays close to the path of gradient flow if the learning rate is small and finite, but on a modified loss. This modified loss is composed of the original loss function and an implicit regularizer, which penalizes the norms of the minibatch gradients. Under mild assumptions, when the batch size is small the scale of the implicit regularization term is proportional to the ratio of the learning rate to the batch size. We verify empirically that explicitly including the implicit regularizer in the loss can enhance the test accuracy when the learning rate is small.

Share or Not? Learning to Schedule Language-Specific Capacity for Multilingual Translation

- Biao Zhang, Ankur Bapna, Rico Sennrich, Orhan Firat
- abstract@open-review(Oral): Using a mix of shared and language-specific (LS) parameters has shown promise in multilingual neural machine translation (MNMT), but the question of when and where LS capacity matters most is still under-studied. We offer such a study by proposing conditional language-specific routing (CLSR). CLSR employs hard binary gates conditioned on token representations to dynamically select LS or shared paths. By manipulating these gates, it can schedule LS capacity across sub-layers in MNMT subject to the guidance of translation signals and budget constraints. Moreover, CLSR can easily scale up to massively multilingual settings. Experiments with Transformer on OPUS-100 and WMT datasets show that: 1) MNMT is sensitive to both the amount and the position of LS modeling: distributing 10%-30% LS computation to the top and/or bottom encoder/decoder layers delivers the best performance; and 2) one-to-many translation benefits more from CLSR compared to many-to-one translation, particularly with unbalanced training data. Our study further verifies the trade-off between the shared capacity and LS capacity for multilingual translation. We corroborate our analysis by confirming the soundness of our findings as foundation of our improved multilingual Transformers. Source code and models are available at https://github.com/bzhangGo/zero/tree/iclr2021_clsr.

Transient Non-stationarity and Generalisation in Deep Reinforcement Learning

- Maximilian Igl, Gregory Farquhar, Jelena Luketina, Wendelin Boehmer, Shimon Whiteson
- abstract@open-review(Poster): Non-stationarity can arise in Reinforcement Learning (RL) even in stationary environments. For example, most RL algorithms collect new data throughout training, using a non-stationary behaviour policy. Due to the transience of this non-stationarity, it is often not explicitly addressed in deep RL and a single neural network is continually updated. However, we find evidence that neural networks exhibit a memory effect, where these transient non-stationarities can permanently impact the latent representation and adversely affect generalisation performance. Consequently, to improve generalisation of deep RL agents, we propose Iterated Relearning (ITER). ITER augments standard RL training by repeated knowledge transfer of the current policy into a freshly initialised network, which thereby experiences less non-stationarity during training. Experimentally, we show that ITER improves performance on the challenging generalisation benchmarks ProcGen and Multiroom.

Lossless Compression of Structured Convolutional Models via Lifting

- Gustav Sourek, Filip Zelezny, Ondrej Kuzelka
- abstract@open-review(Poster): Lifting is an efficient technique to scale up graphical models generalized to relational domains by exploiting the underlying symmetries. Concurrently, neural models are continuously expanding from grid-like tensor data into structured representations, such as various attributed graphs and relational databases. To address the irregular structure of the data, the models typically extrapolate on the idea of convolution, effectively introducing parameter sharing in their, dynamically unfolded, computation graphs. The computation graphs themselves then reflect the

symmetries of the underlying data, similarly to the lifted graphical models. Inspired by lifting, we introduce a simple and efficient technique to detect the symmetries and compress the neural models without loss of any information. We demonstrate through experiments that such compression can lead to significant speedups of structured convolutional models, such as various Graph Neural Networks, across various tasks, such as molecule classification and knowledge-base completion.

Analyzing the Expressive Power of Graph Neural Networks in a Spectral Perspective

- Muhammet Balcilar, Guillaume Renton, Pierre Héroux, Benoit Gaüzère, Sébastien Adam, Paul Honeine
- abstract@open-review(Poster): In the recent literature of Graph Neural Networks (GNN), the expressive power of models has been studied through their capability to distinguish if two given graphs are isomorphic or not. Since the graph isomorphism problem is NP-intermediate, and Weisfeiler-Lehman (WL) test can give sufficient but not enough evidence in polynomial time, the theoretical power of GNNs is usually evaluated by the equivalence of WL-test order, followed by an empirical analysis of the models on some reference inductive and transductive datasets. However, such analysis does not account the signal processing pipeline, whose capability is generally evaluated in the spectral domain. In this paper, we argue that a spectral analysis of GNNs behavior can provide a complementary point of view to go one step further in the understanding of GNNs. By bridging the gap between the spectral and spatial design of graph convolutions, we theoretically demonstrate some equivalence of the graph convolution process regardless it is designed in the spatial or the spectral domain. Using this connection, we managed to re-formulate most of the state-of-the-art graph neural networks into one common framework. This general framework allows to lead a spectral analysis of the most popular GNNs, explaining their performance and showing their limits according to spectral point of view. Our theoretical spectral analysis is confirmed by experiments on various graph databases. Furthermore, we demonstrate the necessity of high and/or band-pass filters on a graph dataset, while the majority of GNN is limited to only low-pass and inevitably it fails.

End-to-end Adversarial Text-to-Speech

- Jeff Donahue, Sander Dieleman, Mikolaj Binkowski, Erich Elsen, Karen Simonyan
- abstract@open-review(Oral): Modern text-to-speech synthesis pipelines typically involve multiple processing stages, each of which is designed or learnt independently from the rest. In this work, we take on the challenging task of learning to synthesise speech from normalised text or phonemes in an end-to-end manner, resulting in models which operate directly on character or phoneme input sequences and produce raw speech audio outputs. Our proposed generator is feed-forward and thus efficient for both training and inference, using a differentiable alignment scheme based on token length prediction. It learns to produce high fidelity audio through a combination of adversarial feedback and prediction losses constraining the generated audio to roughly match the ground truth in terms of its total duration and mel-spectrogram. To allow the model to capture temporal variation in the generated audio, we employ soft dynamic time warping in the spectrogram-based prediction loss. The resulting model achieves a mean opinion score exceeding 4 on a 5 point scale, which is comparable to the state-of-the-art models relying on multi-stage training and additional supervision.

A unifying view on implicit bias in training linear neural networks

- Chulhee Yun, Shankar Krishnan, Hossein Mobahi
- abstract@open-review(Poster): We study the implicit bias of gradient flow (i.e., gradient descent with infinitesimal step size) on linear neural network training. We propose a tensor formulation of neural networks that includes fully-connected, diagonal, and convolutional networks as special cases, and investigate the linear version of the formulation called linear tensor networks. With this formulation, we can characterize the convergence direction of the network parameters as singular vectors of a tensor defined by the network. For L-layer linear tensor networks that are orthogonally decomposable, we show that gradient flow on separable classification finds a stationary point of the $\ell_{2/L}$ max-margin problem in a "transformed" input space defined by the network. For underdetermined regression, we prove that gradient flow finds a global minimum which minimizes a norm-like function that interpolates between weighted ℓ_1 and ℓ_2 norms in the transformed input space. Our theorems subsume existing results in the literature while removing standard convergence assumptions. We also provide experiments that corroborate our analysis.

Balancing Constraints and Rewards with Meta-Gradient D4PG

- Dan A. Calian, Daniel J Mankowitz, Tom Zahavy, Zhongwen Xu, Junhyuk Oh, Nir Levine, Timothy Mann
- abstract@open-review(Poster): Deploying Reinforcement Learning (RL) agents to solve real-world applications often requires satisfying complex system constraints. Often the constraint thresholds are incorrectly set due to the complex nature of a system or the inability to verify the thresholds offline (e.g, no simulator or reasonable offline evaluation procedure exists). This results in solutions where a task cannot be solved without violating the constraints. However, in many real-world cases, constraint violations are undesirable yet they are not catastrophic, motivating the need for soft-constrained RL approaches. We present two soft-constrained RL approaches that utilize meta-gradients to find a good trade-off between expected return and minimizing constraint violations. We demonstrate the effectiveness of these approaches by showing that they consistently outperform the baselines across four different Mujoco domains.

Robust Curriculum Learning: from clean label detection to noisy label self-correction

- Tianyi Zhou, Shengjie Wang, Jeff Bilmes
- abstract@open-review(Poster): Neural network training can easily overfit noisy labels resulting in poor generalization performance. Existing methods address this problem by (1) filtering out the noisy data and only using the clean data for training or (2) relabeling the noisy data by the model during training or by another model trained only on a clean dataset. However, the former does not leverage the features' information of wrongly-labeled data, while the latter may produce wrong pseudo-labels for some data and introduce extra noises. In this paper, we propose a smooth transition and interplay between these two strategies as a curriculum that selects training samples dynamically. In particular, we start with learning from clean data and then gradually move to learn noisy-labeled data with pseudo labels produced by a time-ensemble of the model and data augmentations. Instead of using the instantaneous loss computed at the current step, our data selection is based on the dynamics of both the loss and output consistency for each sample across historical steps and different data augmentations, resulting in more precise detection of both clean labels and correct pseudo labels. On multiple benchmarks of noisy labels, we show that our curriculum learning strategy can significantly improve the test accuracy without any auxiliary model or extra clean data.

Clairvoyance: A Pipeline Toolkit for Medical Time Series

- Daniel Jarrett, Jinsung Yoon, Ioana Bica, Zhaozhi Qian, Ari Ercole, Mihaela van der Schaar
- abstract@open-review(Poster): Time-series learning is the bread and butter of data-driven *clinical decision support*, and the recent explosion in ML research has demonstrated great potential in various healthcare settings. At the same time, medical time-series problems in the wild are challenging due to their highly *composite* nature: They entail design choices and interactions among components that preprocess data, impute missing values, select features, issue predictions, estimate uncertainty, and interpret models. Despite exponential growth in electronic patient data, there is a remarkable gap between the potential and realized utilization of ML for clinical research and decision support. In particular, orchestrating a real-world project lifecycle poses challenges in engineering (i.e. hard to build), evaluation (i.e. hard to assess), and efficiency (i.e. hard to optimize). Designed to address these issues simultaneously, Clairvoyance proposes a unified, end-to-end, autoML-friendly pipeline that serves as a (i) software toolkit, (ii) empirical standard, and (iii) interface for optimization. Our ultimate goal lies in facilitating transparent and reproducible experimentation with complex inference workflows, providing integrated pathways for (1) personalized prediction, (2) treatment-effect estimation, and (3) information acquisition. Through illustrative examples on real-world data in outpatient, general wards, and intensive-care settings, we illustrate the applicability of the pipeline paradigm on core tasks

in the healthcare journey. To the best of our knowledge, Clairvoyance is the first to demonstrate viability of a comprehensive and automatable pipeline for clinical time-series ML.

Plan-Based Relaxed Reward Shaping for Goal-Directed Tasks

- Ingmar Schubert, Ozgur S Oguz, Marc Toussaint
- abstract@open-review(Poster): In high-dimensional state spaces, the usefulness of Reinforcement Learning (RL) is limited by the problem of exploration. This issue has been addressed using potential-based reward shaping (PB-RS) previously. In the present work, we introduce Final-Volume-Preserving Reward Shaping (FV-RS). FV-RS relaxes the strict optimality guarantees of PB-RS to a guarantee of preserved long-term behavior. Being less restrictive, FV-RS allows for reward shaping functions that are even better suited for improving the sample efficiency of RL algorithms. In particular, we consider settings in which the agent has access to an approximate plan. Here, we use examples of simulated robotic manipulation tasks to demonstrate that plan-based FV-RS can indeed significantly improve the sample efficiency of RL over plan-based PB-RS.

Improving VAEs' Robustness to Adversarial Attack

- Matthew JF Willetts, Alexander Camuto, Tom Rainforth, S Roberts, Christopher C Holmes
- abstract@open-review(Poster): Variational autoencoders (VAEs) have recently been shown to be vulnerable to adversarial attacks, wherein they are fooled into reconstructing a chosen target image. However, how to defend against such attacks remains an open problem. We make significant advances in addressing this issue by introducing methods for producing adversarially robust VAEs. Namely, we first demonstrate that methods proposed to obtain disentangled latent representations produce VAEs that are more robust to these attacks. However, this robustness comes at the cost of reducing the quality of the reconstructions. We ameliorate this by applying disentangling methods to hierarchical VAEs. The resulting models produce high--fidelity autoencoders that are also adversarially robust. We confirm their capabilities on several different datasets and with current state-of-the-art VAE adversarial attacks, and also show that they increase the robustness of downstream tasks to attack.

Gauge Equivariant Mesh CNNs: Anisotropic convolutions on geometric graphs

- Pim De Haan, Maurice Weiler, Taco Cohen, Max Welling
- abstract@open-review(Spotlight): A common approach to define convolutions on meshes is to interpret them as a graph and apply graph convolutional networks (GCNs). Such GCNs utilize isotropic kernels and are therefore insensitive to the relative orientation of vertices and thus to the geometry of the mesh as a whole. We propose Gauge Equivariant Mesh CNNs which generalize GCNs to apply anisotropic gauge equivariant kernels. Since the resulting features carry orientation information, we introduce a geometric message passing scheme defined by parallel transporting features over mesh edges. Our experiments validate the significantly improved expressivity of the proposed model over conventional GCNs and other methods.

Differentiable Segmentation of Sequences

- Erik Scharwächter, Jonathan Lennartz, Emmanuel Müller
- abstract@open-review(Poster): Segmented models are widely used to describe non-stationary sequential data with discrete change points. Their estimation usually requires solving a mixed discrete-continuous optimization problem, where the segmentation is the discrete part and all other model parameters are continuous. A number of estimation algorithms have been developed that are highly specialized for their specific model assumptions. The dependence on non-standard algorithms makes it hard to integrate segmented models in state-of-the-art deep learning architectures that critically depend on gradient-based optimization techniques. In this work, we formulate a relaxed variant of segmented models that enables joint estimation of all model parameters, including the segmentation, with gradient descent. We build on recent advances in learning continuous warping functions and propose a novel family of warping functions based on the two-sided power (TSP) distribution. TSP-based warping functions are differentiable, have simple closed-form expressions, and can represent segmentation functions exactly. Our formulation includes the important class of segmented generalized linear models as a special case, which makes it highly versatile. We use our approach to model the spread of COVID-19 with Poisson regression, apply it on a change point detection task, and learn classification models with concept drift. The experiments show that our approach effectively learns all these tasks with standard algorithms for gradient descent.

Generalization bounds via distillation

- Daniel Hsu, Ziwei Ji, Matus Telgarsky, Lan Wang
- abstract@open-review(Spotlight): This paper theoretically investigates the following empirical phenomenon: given a high-complexity network with poor generalization bounds, one can distill it into a network with nearly identical predictions but low complexity and vastly smaller generalization bounds. The main contribution is an analysis showing that the original network inherits this good generalization bound from its distillation, assuming the use of well-behaved data augmentation. This bound is presented both in an abstract and in a concrete form, the latter complemented by a reduction technique to handle modern computation graphs featuring convolutional layers, fully-connected layers, and skip connections, to name a few. To round out the story, a (looser) classical uniform convergence analysis of compression is also presented, as well as a variety of experiments on cifar and mnist demonstrating similar generalization performance between the original network and its distillation.

Learning Mesh-Based Simulation with Graph Networks

- Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, Peter Battaglia
- abstract@open-review(Spotlight): Mesh-based simulations are central to modeling complex physical systems in many disciplines across science and engineering. Mesh representations support powerful numerical integration methods and their resolution can be adapted to strike favorable trade-offs between accuracy and efficiency. However, high-dimensional scientific simulations are very expensive to run, and solvers and parameters must often be tuned individually to each system studied. Here we introduce MeshGraphNets, a framework for learning mesh-based simulations using graph neural networks. Our model can be trained to pass messages on a mesh graph and to adapt the mesh discretization during forward simulation. Our results show it can accurately predict the dynamics of a wide range of physical systems, including aerodynamics, structural mechanics, and cloth. The model's adaptivity supports learning resolution-independent dynamics and can scale to more complex state spaces at test time. Our method is also highly efficient, running 1-2 orders of magnitude faster than the simulation on which it is trained. Our approach broadens the range of problems on which neural network simulators can operate and promises to improve the efficiency of complex, scientific modeling tasks.

GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing

- Tao Yu, Chien-Sheng Wu, Xi Victoria Lin, bailin wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, richard socher, Caiming Xiong
- abstract@open-review(Poster): We present GraPPa, an effective pre-training approach for table semantic parsing that learns a compositional inductive bias in the joint representations of textual and tabular data. We construct synthetic question-SQL pairs over high-quality tables via a synchronous context-free grammar (SCFG). We pre-train our model on the synthetic data to inject important structural properties commonly found in semantic parsing into the pre-training language model. To maintain the model's ability to represent real-world data, we also include masked language modeling (MLM) on several existing table-related datasets to regularize our pre-training process. Our proposed pre-training strategy is much data-efficient. When incorporated with

strong base semantic parsers, GraPPa achieves new state-of-the-art results on four popular fully supervised and weakly supervised table semantic parsing tasks.

Sliced Kernelized Stein Discrepancy

- Wenbo Gong, Yingzhen Li, José Miguel Hernández-Lobato
- abstract@open-review(Poster): Kernelized Stein discrepancy (KSD), though being extensively used in goodness-of-fit tests and model learning, suffers from the curse-of-dimensionality. We address this issue by proposing the sliced Stein discrepancy and its scalable and kernelized variants, which employs kernel-based test functions defined on the optimal one-dimensional projections. When applied to goodness-of-fit tests, extensive experiments show the proposed discrepancy significantly outperforms KSD and various baselines in high dimensions. For model learning, we show its advantages by training an independent component analysis when compared with existing Stein discrepancy baselines. We further propose a novel particle inference method called sliced Stein variational gradient descent (S-SVGD) which alleviates the mode-collapse issue of SVGD in training variational autoencoders.

Variational Intrinsic Control Revisited

- Taehwan Kwon
- abstract@open-review(Poster): In this paper, we revisit variational intrinsic control (VIC), an unsupervised reinforcement learning method for finding the largest set of intrinsic options available to an agent. In the original work by Gregor et al. (2016), two VIC algorithms were proposed: one that represents the options explicitly, and the other that does it implicitly. We show that the intrinsic reward used in the latter is subject to bias in stochastic environments, causing convergence to suboptimal solutions. To correct this behavior, we propose two methods respectively based on the transitional probability model and Gaussian Mixture Model. We substantiate our claims through rigorous mathematical derivations and experimental analyses.

On Statistical Bias In Active Learning: How and When to Fix It

- Sebastian Farquhar, Yarin Gal, Tom Rainforth
- abstract@open-review(Spotlight): Active learning is a powerful tool when labelling data is expensive, but it introduces a bias because the training data no longer follows the population distribution. We formalize this bias and investigate the situations in which it can be harmful and sometimes even helpful. We further introduce novel corrective weights to remove bias when doing so is beneficial. Through this, our work not only provides a useful mechanism that can improve the active learning approach, but also an explanation for the empirical successes of various existing approaches which ignore this bias. In particular, we show that this bias can be actively helpful when training overparameterized models---like neural networks---with relatively modest dataset sizes.

Winning the L2RPN Challenge: Power Grid Management via Semi-Markov Afterstate Actor-Critic

- Deunsol Yoon, Sunghoon Hong, Byung-Jun Lee, Kee-Eung Kim
- abstract@open-review(Spotlight): Safe and reliable electricity transmission in power grids is crucial for modern society. It is thus quite natural that there has been a growing interest in the automatic management of power grids, exemplified by the Learning to Run a Power Network Challenge (L2RPN), modeling the problem as a reinforcement learning (RL) task. However, it is highly challenging to manage a real-world scale power grid, mostly due to the massive scale of its state and action space. In this paper, we present an off-policy actor-critic approach that effectively tackles the unique challenges in power grid management by RL, adopting the hierarchical policy together with the afterstate representation. Our agent ranked first in the latest challenge (L2RPN WCCI 2020), being able to avoid disastrous situations while maintaining the highest level of operational efficiency in every test scenarios. This paper provides a formal description of the algorithmic aspect of our approach, as well as further experimental studies on diverse power grids.

HyperDynamics: Meta-Learning Object and Agent Dynamics with Hypernetworks

- Zhou Xian, Shamit Lal, Hsiao-Yu Tung, Emmanouil Antonios Platanios, Katerina Fragkiadaki
- abstract@open-review(Poster): We propose HyperDynamics, a dynamics meta-learning framework that conditions on an agent's interactions with the environment and optionally its visual observations, and generates the parameters of neural dynamics models based on inferred properties of the dynamical system. Physical and visual properties of the environment that are not part of the low-dimensional state yet affect its temporal dynamics are inferred from the interaction history and visual observations, and are implicitly captured in the generated parameters. We test HyperDynamics on a set of object pushing and locomotion tasks. It outperforms existing dynamics models in the literature that adapt to environment variations by learning dynamics over high dimensional visual observations, capturing the interactions of the agent in recurrent state representations, or using gradient-based meta-optimization. We also show our method matches the performance of an ensemble of separately trained experts, while also being able to generalize well to unseen environment variations at test time. We attribute its good performance to the multiplicative interactions between the inferred system properties—captured in the generated parameters—and the low-dimensional state representation of the dynamical system.

Towards Resolving the Implicit Bias of Gradient Descent for Matrix Factorization: Greedy Low-Rank Learning

- Zhiyuan Li, Yuping Luo, Kaifeng Lyu
- abstract@open-review(Poster): Matrix factorization is a simple and natural test-bed to investigate the implicit regularization of gradient descent. Gunasekar et al. (2017) conjectured that gradient flow with infinitesimal initialization converges to the solution that minimizes the nuclear norm, but a series of recent papers argued that the language of norm minimization is not sufficient to give a full characterization for the implicit regularization. In this work, we provide theoretical and empirical evidence that for depth-2 matrix factorization, gradient flow with infinitesimal initialization is mathematically equivalent to a simple heuristic rank minimization algorithm, Greedy Low-Rank Learning, under some reasonable assumptions. This generalizes the rank minimization view from previous works to a much broader setting and enables us to construct counter-examples to refute the conjecture from Gunasekar et al. (2017). We also extend the results to the case where depth >= 3, and we show that the benefit of being deeper is that the above convergence has a much weaker dependence over initialization magnitude so that this rank minimization is more likely to take effect for initialization with practical scale.

Private Post-GAN Boosting

- Marcel Neunhoeffer, Steven Wu, Cynthia Dwork
- abstract@open-review(Poster): Differentially private GANs have proven to be a promising approach for generating realistic synthetic data without compromising the privacy of individuals. Due to the privacy-protective noise introduced in the training, the convergence of GANs becomes even more elusive, which often leads to poor utility in the output generator at the end of training. We propose Private post-GAN boosting (Private PGB), a differentially private method that combines samples produced by the sequence of generators obtained during GAN training to create a high-quality synthetic dataset. To that end, our method leverages the Private Multiplicative Weights method (Hardt and Rothblum, 2010) to reweight generated samples. We evaluate Private PGB on two dimensional toy data, MNIST images, US Census data and a standard machine learning prediction task. Our experiments show that Private PGB improves upon a standard private GAN approach across a collection of quality measures. We also provide a non-private variant of PGB that improves the data quality of standard GAN training.

Characterizing signal propagation to close the performance gap in unnormalized ResNets

- Andrew Brock, Soham De, Samuel L Smith
- abstract@open-review(Poster): Batch Normalization is a key component in almost all state-of-the-art image classifiers, but it also introduces practical challenges: it breaks the independence between training examples within a batch, can incur compute and memory overhead, and often results in unexpected bugs. Building on recent theoretical analyses of deep ResNets at initialization, we propose a simple set of analysis tools to characterize signal propagation on the forward pass, and leverage these tools to design highly performant ResNets without activation normalization layers. Crucial to our success is an adapted version of the recently proposed Weight Standardization. Our analysis tools show how this technique preserves the signal in ReLU networks by ensuring that the per-channel activation means do not grow with depth. Across a range of FLOP budgets, our networks attain performance competitive with state-of-the-art EfficientNets on ImageNet.

Prototypical Contrastive Learning of Unsupervised Representations

- Junnan Li, Pan Zhou, Caiming Xiong, Steven Hoi
- abstract@open-review(Poster): This paper presents Prototypical Contrastive Learning (PCL), an unsupervised representation learning method that bridges contrastive learning with clustering. PCL not only learns low-level features for the task of instance discrimination, but more importantly, it implicitly encodes semantic structures of the data into the learned embedding space. Specifically, we introduce prototypes as latent variables to help find the maximum-likelihood estimation of the network parameters in an Expectation-Maximization framework. We iteratively perform E-step as finding the distribution of prototypes via clustering and M-step as optimizing the network via contrastive learning. We propose ProtoNCE loss, a generalized version of the InfoNCE loss for contrastive learning, which encourages representations to be closer to their assigned prototypes. PCL outperforms state-of-the-art instance-wise contrastive learning methods on multiple benchmarks with substantial improvement in low-resource transfer learning. Code and pretrained models are available at https://github.com/salesforce/PCL.

Hyperbolic Neural Networks++

- Ryohei Shimizu, YUSUKE Mukuta, Tatsuya Harada
- abstract@open-review(Poster): Hyperbolic spaces, which have the capacity to embed tree structures without distortion owing to their exponential volume growth, have recently been applied to machine learning to better capture the hierarchical nature of data. In this study, we generalize the fundamental components of neural networks in a single hyperbolic geometry model, namely, the Poincaré ball model. This novel methodology constructs a multinomial logistic regression, fully-connected layers, convolutional layers, and attention mechanisms under a unified mathematical interpretation, without increasing the parameters. Experiments show the superior parameter efficiency of our methods compared to conventional hyperbolic components, and stability and outperformance over their Euclidean counterparts.

Lipschitz Recurrent Neural Networks

- N. Benjamin Erichson, Omri Azencot, Alejandro Queiruga, Liam Hodgkinson, Michael W. Mahoney
- abstract@open-review(Poster): Viewing recurrent neural networks (RNNs) as continuous-time dynamical systems, we propose a recurrent unit that describes the hidden state's evolution with two parts: a well-understood linear component plus a Lipschitz nonlinearity. This particular functional form facilitates stability analysis of the long-term behavior of the recurrent unit using tools from nonlinear systems theory. In turn, this enables architectural design decisions before experimentation. Sufficient conditions for global stability of the recurrent unit are obtained, motivating a novel scheme for constructing hidden-to-hidden matrices. Our experiments demonstrate that the Lipschitz RNN can outperform existing recurrent units on a range of benchmark tasks, including computer vision, language modeling and speech prediction tasks. Finally, through Hessian-based analysis we demonstrate that our Lipschitz recurrent unit is more robust with respect to input and parameter perturbations as compared to other continuous-time RNNs.

A statistical theory of cold posteriors in deep neural networks

- Laurence Aitchison
- abstract@open-review(Poster): To get Bayesian neural networks to perform comparably to standard neural networks it is usually necessary to artificially reduce uncertainty using a tempered or cold posterior. This is extremely concerning: if the prior is accurate, Bayes inference/decision theory is optimal, and any artificial changes to the posterior should harm performance. While this suggests that the prior may be at fault, here we argue that in fact, BNNs for image classification use the wrong likelihood. In particular, standard image benchmark datasets such as CIFAR-10 are carefully curated. We develop a generative model describing curation which gives a principled Bayesian account of cold posteriors, because the likelihood under this new generative model closely matches the tempered likelihoods used in past work.

Boost then Convolve: Gradient Boosting Meets Graph Neural Networks

- Sergei Ivanov, Liudmila Prokhorenkova
- abstract@open-review(Poster): Graph neural networks (GNNs) are powerful models that have been successful in various graph representation learning tasks. Whereas gradient boosted decision trees (GBDT) often outperform other machine learning methods when faced with heterogeneous tabular data. But what approach should be used for graphs with tabular node features? Previous GNN models have mostly focused on networks with homogeneous sparse features and, as we show, are suboptimal in the heterogeneous setting. In this work, we propose a novel architecture that trains GBDT and GNN jointly to get the best of both worlds: the GBDT model deals with heterogeneous features, while GNN accounts for the graph structure. Our model benefits from end-to-end optimization by allowing new trees to fit the gradient updates of GNN. With an extensive experimental comparison to the leading GBDT and GNN models, we demonstrate a significant increase in performance on a variety of graphs with tabular features. The code is available: https://github.com/nd7141/bgnn.

Genetic Soft Updates for Policy Evolution in Deep Reinforcement Learning

- Enrico Marchesini, Davide Corsi, Alessandro Farinelli
- abstract@open-review(Poster): The combination of Evolutionary Algorithms (EAs) and Deep Reinforcement Learning (DRL) has been recently proposed to merge the benefits of both solutions. Existing mixed approaches, however, have been successfully applied only to actor-critic methods and present significant overhead. We address these issues by introducing a novel mixed framework that exploits a periodical genetic evaluation to soft update the weights of a DRL agent. The resulting approach is applicable with any DRL method and, in a worst-case scenario, it does not exhibit detrimental behaviours. Experiments in robotic applications and continuous control benchmarks demonstrate the versatility of our approach that significantly outperforms prior DRL, EAs, and mixed approaches. Finally, we employ formal verification to confirm the policy improvement, mitigating the inefficient exploration and hyper-parameter sensitivity of DRL.ment, mitigating the inefficient exploration and hyper-parameter sensitivity of DRL.

Spatially Structured Recurrent Modules

- Nasim Rahaman, Anirudh Goyal, Muhammad Waleed Gondal, Manuel Wuthrich, Stefan Bauer, Yash Sharma, Yoshua Bengio, Bernhard Schölkopf
- abstract@open-review(Poster): Capturing the structure of a data-generating process by means of appropriate inductive biases can help in learning models that generalise well and are robust to changes in the input distribution. While methods that harness spatial and temporal structures find broad application, recent work has demonstrated the potential of models that leverage sparse and modular structure using an ensemble of sparingly interacting modules. In

this work, we take a step towards dynamic models that are capable of simultaneously exploiting both modular and spatiotemporal structures. To this end, we model the dynamical system as a collection of autonomous but sparsely interacting sub-systems that interact according to a learned topology which is informed by the spatial structure of the underlying system. This gives rise to a class of models that are well suited for capturing the dynamics of systems that only offer local views into their state, along with corresponding spatial locations of those views. On the tasks of video prediction from cropped frames and multi-agent world modelling from partial observations in the challenging Starcraft2 domain, we find our models to be more robust to the number of available views and better capable of generalisation to novel tasks without additional training than strong baselines that perform equally well or better on the training distribution.

Expressive Power of Invariant and Equivariant Graph Neural Networks

- Waiss Azizian, marc lelarge
- abstract@open-review(Spotlight): Various classes of Graph Neural Networks (GNN) have been proposed and shown to be successful in a wide range of applications with graph structured data. In this paper, we propose a theoretical framework able to compare the expressive power of these GNN architectures. The current universality theorems only apply to intractable classes of GNNs. Here, we prove the first approximation guarantees for practical GNNs, paving the way for a better understanding of their generalization. Our theoretical results are proved for invariant GNNs computing a graph embedding (permutation of the nodes of the input graph does not affect the output) and equivariant GNNs computing an embedding of the nodes (permutation of the input permutes the output). We show that Folklore Graph Neural Networks (FGNN), which are tensor based GNNs augmented with matrix multiplication are the most expressive architectures proposed so far for a given tensor order. We illustrate our results on the Quadratic Assignment Problem (a NP-Hard combinatorial problem) by showing that FGNNs are able to learn how to solve the problem, leading to much better average performances than existing algorithms (based on spectral, SDP or other GNNs architectures). On a practical side, we also implement masked tensors to handle batches of graphs of varying sizes.

On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines

- Marius Mosbach, Maksym Andriushchenko, Dietrich Klakow
- abstract@open-review(Poster): Fine-tuning pre-trained transformer-based language models such as BERT has become a common practice dominating leaderboards across various NLP benchmarks. Despite the strong empirical performance of fine-tuned models, fine-tuning is an unstable process: training the same model with multiple random seeds can result in a large variance of the task performance. Previous literature (Devlin et al., 2019; Lee et al., 2020; Dodge et al., 2020) identified two potential reasons for the observed instability: catastrophic forgetting and small size of the fine-tuning datasets. In this paper, we show that both hypotheses fail to explain the fine-tuning instability. We analyze BERT, RoBERTa, and ALBERT, fine-tuned on commonly used datasets from the GLUE benchmark, and show that the observed instability is caused by optimization difficulties that lead to vanishing gradients. Additionally, we show that the remaining variance of the downstream task performance can be attributed to differences in generalization where fine-tuned models with the same training loss exhibit noticeably different test performance. Based on our analysis, we present a simple but strong baseline that makes fine-tuning BERT-based models significantly more stable than the previously proposed approaches. Code to reproduce our results is available online: https://github.com/uds-lsv/bert-stable-fine-tuning.

End-to-End Egospheric Spatial Memory

- Daniel James Lenton, Stephen James, Ronald Clark, Andrew Davison
- abstract@open-review(Poster): Spatial memory, or the ability to remember and recall specific locations and objects, is central to autonomous agents' ability to carry out tasks in real environments. However, most existing artificial memory modules are not very adept at storing spatial information. We propose a parameter-free module, Egospheric Spatial Memory (ESM), which encodes the memory in an ego-sphere around the agent, enabling expressive 3D representations. ESM can be trained end-to-end via either imitation or reinforcement learning, and improves both training efficiency and final performance against other memory baselines on both drone and manipulator visuomotor control tasks. The explicit egocentric geometry also enables us to seamlessly combine the learned controller with other non-learned modalities, such as local obstacle avoidance. We further show applications to semantic segmentation on the ScanNet dataset, where ESM naturally combines image-level and map-level inference modalities. Through our broad set of experiments, we show that ESM provides a general computation graph for embodied spatial reasoning, and the module forms a bridge between real-time mapping systems and differentiable memory architectures. Implementation at: https://github.com/ivy-dl/memory.

LEAF: A Learnable Frontend for Audio Classification

- Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, Marco Tagliasacchi
- abstract@open-review(Poster): Mel-filterbanks are fixed, engineered audio features which emulate human perception and have been used through the history of audio understanding up to today. However, their undeniable qualities are counterbalanced by the fundamental limitations of handmade representations. In this work we show that we can train a single learnable frontend that outperforms mel-filterbanks on a wide range of audio signals, including speech, music, audio events and animal sounds, providing a general-purpose learned frontend for audio classification. To do so, we introduce a new principled, lightweight, fully learnable architecture that can be used as a drop-in replacement of mel-filterbanks. Our system learns all operations of audio features extraction, from filtering to pooling, compression and normalization, and can be integrated into any neural network at a negligible parameter cost. We perform multi-task training on eight diverse audio classification tasks, and show consistent improvements of our model over mel-filterbanks and previous learnable alternatives. Moreover, our system outperforms the current state-of-the-art learnable frontend on Audioset, with orders of magnitude fewer parameters.

Simple Augmentation Goes a Long Way: ADRL for DNN Quantization

- Lin Ning, Guoyang Chen, Weifeng Zhang, Xipeng Shen
- abstract@open-review(Poster): Mixed precision quantization improves DNN performance by assigning different layers with different bit-width values. Searching for the optimal bit-width for each layer, however, remains a challenge. Deep Reinforcement Learning (DRL) shows some recent promise. It however suffers instability due to function approximation errors, causing large variances in the early training stages, slow convergence, and suboptimal policies in the mixed-precision quantization problem. This paper proposes augmented DRL (ADRL) as a way to alleviate these issues. This new strategy augments the neural networks in DRL with a complementary scheme to boost the performance of learning. The paper examines the effectiveness of ADRL both analytically and empirically, showing that it can produce more accurate quantized models than the state of the art DRL-based quantization while improving the learning speed by 4.5-64 times.

The inductive bias of ReLU networks on orthogonally separable data

- Mary Phuong, Christoph H Lampert
- abstract@open-review(Poster): We study the inductive bias of two-layer ReLU networks trained by gradient flow. We identify a class of easy-to-learn (`orthogonally separable') datasets, and characterise the solution that ReLU networks trained on such datasets converge to. Irrespective of network width, the solution turns out to be a combination of two max-margin classifiers: one corresponding to the positive data subset and one corresponding to the negative data subset. The proof is based on the recently introduced concept of extremal sectors, for which we prove a number of properties in the context of orthogonal separability. In particular, we prove stationarity of activation patterns from some time T onwards, which enables a reduction of the ReLU network to an ensemble of linear subnetworks.

Monte-Carlo Planning and Learning with Language Action Value Estimates

- Youngsoo Jang, Seokin Seo, Jongmin Lee, Kee-Eung Kim
- abstract@open-review(Poster): Interactive Fiction (IF) games provide a useful testbed for language-based reinforcement learning agents, posing significant challenges of natural language understanding, commonsense reasoning, and non-myopic planning in the combinatorial search space. Agents based on standard planning algorithms struggle to play IF games due to the massive search space of language actions. Thus, language-grounded planning is a key ability of such agents, since inferring the consequence of language action based on semantic understanding can drastically improve search. In this paper, we introduce Monte-Carlo planning with Language Action Value Estimates (MC-LAVE) that combines a Monte-Carlo tree search with language-driven exploration. MC-LAVE invests more search effort into semantically promising language actions using locally optimistic language value estimates, yielding a significant reduction in the effective search space of language actions. We then present a reinforcement learning approach via MC-LAVE, which alternates between MC-LAVE planning and supervised learning of the self-generated language actions. In the experiments, we demonstrate that our method achieves new high scores in various IF games.

Learning Energy-Based Models by Diffusion Recovery Likelihood

- Ruiqi Gao, Yang Song, Ben Poole, Ying Nian Wu, Diederik P Kingma
- abstract@open-review(Poster): While energy-based models (EBMs) exhibit a number of desirable properties, training and sampling on high-dimensional datasets remains challenging. Inspired by recent progress on diffusion probabilistic models, we present a diffusion recovery likelihood method to tractably learn and sample from a sequence of EBMs trained on increasingly noisy versions of a dataset. Each EBM is trained with recovery likelihood, which maximizes the conditional probability of the data at a certain noise level given their noisy versions at a higher noise level. Optimizing recovery likelihood is more tractable than marginal likelihood, as sampling from the conditional distributions is much easier than sampling from the marginal distributions. After training, synthesized images can be generated by the sampling process that initializes from Gaussian white noise distribution and progressively samples the conditional distributions at decreasingly lower noise levels. Our method generates high fidelity samples on various image datasets. On unconditional CIFAR-10 our method achieves FID 9.58 and inception score 8.30, superior to the majority of GANs. Moreover, we demonstrate that unlike previous work on EBMs, our long-run MCMC samples from the conditional distributions do not diverge and still represent realistic images, allowing us to accurately estimate the normalized density of data even for high-dimensional datasets. Our implementation is available at \url{https://github.com/ruiqigao/recovery_likelihood}.

Capturing Label Characteristics in VAEs

- Tom Joy, Sebastian Schmon, Philip Torr, Siddharth N, Tom Rainforth
- abstract@open-review(Poster): We present a principled approach to incorporating labels in variational autoencoders (VAEs) that captures the rich characteristic information associated with those labels. While prior work has typically conflated these by learning latent variables that directly correspond to label values, we argue this is contrary to the intended effect of supervision in VAEs—capturing rich label characteristics with the latents. For example, we may want to capture the characteristics of a face that make it look young, rather than just the age of the person. To this end, we develop a novel VAE model, the characteristic capturing VAE (CCVAE), which “reparameterizes” supervision through auxiliary variables and a concomitant variational objective. Through judicious structuring of mappings between latent and auxiliary variables, we show that the CCVAE can effectively learn meaningful representations of the characteristics of interest across a variety of supervision schemes. In particular, we show that the CCVAE allows for more effective and more general interventions to be performed, such as smooth traversals within the characteristics for a given label, diverse conditional generation, and transferring characteristics across datapoints.

Linear Mode Connectivity in Multitask and Continual Learning

- Seyed Iman Mirzadeh, Mehrdad Farajtabar, Dilan Gorur, Razvan Pascanu, Hassan Ghasemzadeh
- abstract@open-review(Poster): Continual (sequential) training and multitask (simultaneous) training are often attempting to solve the same overall objective: to find a solution that performs well on all considered tasks. The main difference is in the training regimes, where continual learning can only have access to one task at a time, which for neural networks typically leads to catastrophic forgetting. That is, the solution found for a subsequent task does not perform well on the previous ones anymore. However, the relationship between the different minima that the two training regimes arrive at is not well understood. What sets them apart? Is there a local structure that could explain the difference in performance achieved by the two different schemes? Motivated by recent work showing that different minima of the same task are typically connected by very simple curves of low error, we investigate whether multitask and continual solutions are similarly connected. We empirically find that indeed such connectivity can be reliably achieved and, more interestingly, it can be done by a linear path, conditioned on having the same initialization for both. We thoroughly analyze this observation and discuss its significance for the continual learning process. Furthermore, we exploit this finding to propose an effective algorithm that constrains the sequentially learned minima to behave as the multitask solution. We show that our method outperforms several state of the art continual learning algorithms on various vision benchmarks.

Computational Separation Between Convolutional and Fully-Connected Networks

- eran malach, Shai Shalev-Shwartz
- abstract@open-review(Poster): Convolutional neural networks (CNN) exhibit unmatched performance in a multitude of computer vision tasks. However, the advantage of using convolutional networks over fully-connected networks is not understood from a theoretical perspective. In this work, we show how convolutional networks can leverage locality in the data, and thus achieve a computational advantage over fully-connected networks. Specifically, we show a class of problems that can be efficiently solved using convolutional networks trained with gradient-descent, but at the same time is hard to learn using a polynomial-size fully-connected network.

Rethinking Embedding Coupling in Pre-trained Language Models

- Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, Sebastian Ruder
- abstract@open-review(Poster): We re-evaluate the standard practice of sharing weights between input and output embeddings in state-of-the-art pre-trained language models. We show that decoupled embeddings provide increased modeling flexibility, allowing us to significantly improve the efficiency of parameter allocation in the input embedding of multilingual models. By reallocating the input embedding parameters in the Transformer layers, we achieve dramatically better performance on standard natural language understanding tasks with the same number of parameters during fine-tuning. We also show that allocating additional capacity to the output embedding provides benefits to the model that persist through the fine-tuning stage even though the output embedding is discarded after pre-training. Our analysis shows that larger output embeddings prevent the model's last layers from overspecializing to the pre-training task and encourage Transformer representations to be more general and more transferable to other tasks and languages. Harnessing these findings, we are able to train models that achieve strong performance on the XTREME benchmark without increasing the number of parameters at the fine-tuning stage.

Co-Mixup: Saliency Guided Joint Mixup with Supermodular Diversity

- JangHyun Kim, Wonho Choo, Hosan Jeong, Hyun Oh Song

- abstract@open-review(Oral): While deep neural networks show great performance on fitting to the training distribution, improving the networks' generalization performance to the test distribution and robustness to the sensitivity to input perturbations still remain as a challenge. Although a number of mixup based augmentation strategies have been proposed to partially address them, it remains unclear as to how to best utilize the supervisory signal within each input data for mixup from the optimization perspective. We propose a new perspective on batch mixup and formulate the optimal construction of a batch of mixup data maximizing the data saliency measure of each individual mixup data and encouraging the supermodular diversity among the constructed mixup data. This leads to a novel discrete optimization problem minimizing the difference between submodular functions. We also propose an efficient modular approximation based iterative submodular minimization algorithm for efficient mixup computation per each minibatch suitable for minibatch based neural network training. Our experiments show the proposed method achieves the state of the art generalization, calibration, and weakly supervised localization results compared to other mixup methods. The source code is available at https://github.com/snu-mllab/Co-Mixup.

Physics-aware, probabilistic model order reduction with guaranteed stability

- Sebastian Kaltenbach, Phaedon Stelios Koutsourelakis
- abstract@open-review(Poster): Given (small amounts of) time-series' data from a high-dimensional, fine-grained, multiscale dynamical system, we propose a generative framework for learning an effective, lower-dimensional, coarse-grained dynamical model that is predictive of the fine-grained system's long-term evolution but also of its behavior under different initial conditions. We target fine-grained models as they arise in physical applications (e.g. molecular dynamics, agent-based models), the dynamics of which are strongly non-stationary but their transition to equilibrium is governed by unknown slow processes which are largely inaccessible by brute-force simulations. Approaches based on domain knowledge heavily rely on physical insight in identifying temporally slow features and fail to enforce the long-term stability of the learned dynamics. On the other hand, purely statistical frameworks lack interpretability and rely on large amounts of expensive simulation data (long and multiple trajectories) as they cannot infuse domain knowledge. The generative framework proposed achieves the aforementioned desiderata by employing a flexible prior on the complex plane for the latent, slow processes, and an intermediate layer of physics-motivated latent variables that reduces reliance on data and imbues inductive bias. In contrast to existing schemes, it does not require the a priori definition of projection operators from the fine-grained description and addresses simultaneously the tasks of dimensionality reduction and model estimation. We demonstrate its efficacy and accuracy in multiscale physical systems of particle dynamics where probabilistic, long-term predictions of phenomena not contained in the training data are produced.

Disentangling 3D Prototypical Networks for Few-Shot Concept Learning

- Mihir Prabhudesai, Shamit Lal, Darshan Patil, Hsiao-Yu Tung, Adam W Harley, Katerina Fragkiadaki
- abstract@open-review(Poster): We present neural architectures that disentangle RGB-D images into objects' shapes and styles and a map of the background scene, and explore their applications for few-shot 3D object detection and few-shot concept classification. Our networks incorporate architectural biases that reflect the image formation process, 3D geometry of the world scene, and shape-style interplay. They are trained end-to-end self-supervised by predicting views in static scenes, alongside a small number of 3D object boxes. Objects and scenes are represented in terms of 3D feature grids in the bottleneck of the network. We show the proposed 3D neural representations are compositional: they can generate novel 3D scene feature maps by mixing object shapes and styles, resizing and adding the resulting object 3D feature maps over background scene feature maps. We show object detectors trained on hallucinated 3D neural scenes generalize better to novel environments. We show classifiers for object categories, color, materials, and spatial relationships trained over the disentangled 3D feature sub-spaces generalize better with dramatically fewer exemplars over the current state-of-the-art, and enable a visual question answering system that uses them as its modules to generalize one-shot to novel objects in the scene.

LiftPool: Bidirectional ConvNet Pooling

- Jiaojiao Zhao, Cees G. M. Snoek
- abstract@open-review(Poster): Pooling is a critical operation in convolutional neural networks for increasing receptive fields and improving robustness to input variations. Most existing pooling operations downsample the feature maps, which is a lossy process. Moreover, they are not invertible: upsampling a downscaled feature map can not recover the lost information in the downsampling. By adopting the philosophy of the classical Lifting Scheme from signal processing, we propose LiftPool for bidirectional pooling layers, including LiftDownPool and LiftUpPool. LiftDownPool decomposes a feature map into various downsized sub-bands, each of which contains information with different frequencies. As the pooling function in LiftDownPool is perfectly invertible, by performing LiftDownPool backward, a corresponding up-pooling layer LiftUpPool is able to generate a refined upsampled feature map using the detail subbands, which is useful for image-to-image translation challenges. Experiments show the proposed methods achieve better results on image classification and semantic segmentation, using various backbones. Moreover, LiftDownPool offers better robustness to input corruptions and perturbations.

Latent Convergent Cross Mapping

- Edward De Brouwer, Adam Arany, Jaak Simm, Yves Moreau
- abstract@open-review(Poster): Discovering causal structures of temporal processes is a major tool of scientific inquiry because it helps us better understand and explain the mechanisms driving a phenomenon of interest, thereby facilitating analysis, reasoning, and synthesis for such systems. However, accurately inferring causal structures within a phenomenon based on observational data only is still an open problem. Indeed, this type of data usually consists in short time series with missing or noisy values for which causal inference is increasingly difficult. In this work, we propose a method to uncover causal relations in chaotic dynamical systems from short, noisy and sporadic time series (that is, incomplete observations at infrequent and irregular intervals) where the classical convergent cross mapping (CCM) fails. Our method works by learning a Neural ODE latent process modeling the state-space dynamics of the time series and by checking the existence of a continuous map between the resulting processes. We provide theoretical analysis and show empirically that Latent-CCM can reliably uncover the true causal pattern, unlike traditional methods.

You Only Need Adversarial Supervision for Semantic Image Synthesis

- Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, Anna Khoreva
- abstract@open-review(Poster): Despite their recent successes, GAN models for semantic image synthesis still suffer from poor image quality when trained with only adversarial supervision. Historically, additionally employing the VGG-based perceptual loss has helped to overcome this issue, significantly improving the synthesis quality, but at the same time limiting the progress of GAN models for semantic image synthesis. In this work, we propose a novel, simplified GAN model, which needs only adversarial supervision to achieve high quality results. We re-design the discriminator as a semantic segmentation network, directly using the given semantic label maps as the ground truth for training. By providing stronger supervision to the discriminator as well as to the generator through spatially- and semantically-aware discriminator feedback, we are able to synthesize images of higher fidelity with better alignment to their input label maps, making the use of the perceptual loss superfluous. Moreover, we enable high-quality multi-modal image synthesis through global and local sampling of a 3D noise tensor injected into the generator, which allows complete or partial image change. We show that images synthesized by our model are more diverse and follow the color and texture distributions of real images more closely. We achieve an average improvement of 6 FID and 5 mIoU points over the state of the art across different datasets using only adversarial supervision.

MARS: Markov Molecular Sampling for Multi-objective Drug Discovery

- Yutong Xie, Chence Shi, Hao Zhou, Yuwei Yang, Weinan Zhang, Yong Yu, Lei Li

- abstract@open-review(Spotlight): Searching for novel molecules with desired chemical properties is crucial in drug discovery. Existing work focuses on developing neural models to generate either molecular sequences or chemical graphs. However, it remains a big challenge to find novel and diverse compounds satisfying several properties. In this paper, we propose MARS, a method for multi-objective drug molecule discovery. MARS is based on the idea of generating the chemical candidates by iteratively editing fragments of molecular graphs. To search for high-quality candidates, it employs Markov chain Monte Carlo sampling (MCMC) on molecules with an annealing scheme and an adaptive proposal. To further improve sample efficiency, MARS uses a graph neural network (GNN) to represent and select candidate edits, where the GNN is trained on-the-fly with samples from MCMC. Experiments show that MARS achieves state-of-the-art performance in various multi-objective settings where molecular bio-activity, drug-likeness, and synthesizability are considered. Remarkably, in the most challenging setting where all four objectives are simultaneously optimized, our approach outperforms previous methods significantly in comprehensive evaluations. The code is available at https://github.com/yutxie/mars.

On Self-Supervised Image Representations for GAN Evaluation

- Stanislav Morozov, Andrey Voynov, Artem Babenko
- abstract@open-review(Spotlight): The embeddings from CNNs pretrained on Imagenet classification are de-facto standard image representations for assessing GANs via FID, Precision and Recall measures. Despite broad previous criticism of their usage for non-Imagenet domains, these embeddings are still the top choice in most of the GAN literature.

In this paper, we advocate the usage of the state-of-the-art self-supervised representations to evaluate GANs on the established non-Imagenet benchmarks. These representations, typically obtained via contrastive learning, are shown to provide better transfer to new tasks and domains, therefore, can serve as more universal embeddings of natural images. With extensive comparison of the recent GANs on the common datasets, we show that self-supervised representations produce a more reasonable ranking of models in terms of FID/Precision/Recall, while the ranking with classification-pretrained embeddings often can be misleading.

A Diffusion Theory For Deep Learning Dynamics: Stochastic Gradient Descent Exponentially Favors Flat Minima

- Zeke Xie, Issei Sato, Masashi Sugiyama
- abstract@open-review(Poster): Stochastic Gradient Descent (SGD) and its variants are mainstream methods for training deep networks in practice. SGD is known to find a flat minimum that often generalizes well. However, it is mathematically unclear how deep learning can select a flat minimum among so many minima. To answer the question quantitatively, we develop a density diffusion theory to reveal how minima selection quantitatively depends on the minima sharpness and the hyperparameters. To the best of our knowledge, we are the first to theoretically and empirically prove that, benefited from the Hessian-dependent covariance of stochastic gradient noise, SGD favors flat minima exponentially more than sharp minima, while Gradient Descent (GD) with injected white noise favors flat minima only polynomially more than sharp minima. We also reveal that either a small learning rate or large-batch training requires exponentially many iterations to escape from minima in terms of the ratio of the batch size and learning rate. Thus, large-batch training cannot search flat minima efficiently in a realistic computational time.

Robust Learning of Fixed-Structure Bayesian Networks in Nearly-Linear Time

- Yu Cheng, Honghao Lin
- abstract@open-review(Poster): We study the problem of learning Bayesian networks where an ϵ-fraction of the samples are adversarially corrupted. We focus on the fully-observable case where the underlying graph structure is known. In this work, we present the first nearly-linear time algorithm for this problem with a dimension-independent error guarantee. Previous robust algorithms with comparable error guarantees are slower by at least a factor of (d/ϵ), where d is the number of variables in the Bayesian network and ϵ is the fraction of corrupted samples.

Our algorithm and analysis are considerably simpler than those in previous work. We achieve this by establishing a direct connection between robust learning of Bayesian networks and robust mean estimation. As a subroutine in our algorithm, we develop a robust mean estimation algorithm whose runtime is nearly-linear in the number of nonzeros in the input samples, which may be of independent interest.

Contrastive Explanations for Reinforcement Learning via Embedded Self Predictions

- Zhengxian Lin, Kin-Ho Lam, Alan Fern
- abstract@open-review(Oral): We investigate a deep reinforcement learning (RL) architecture that supports explaining why a learned agent prefers one action over another. The key idea is to learn action-values that are directly represented via human-understandable properties of expected futures. This is realized via the embedded self-prediction (ESP) model, which learns said properties in terms of human provided features. Action preferences can then be explained by contrasting the future properties predicted for each action. To address cases where there are a large number of features, we develop a novel method for computing minimal sufficient explanations from an ESP. Our case studies in three domains, including a complex strategy game, show that ESP models can be effectively learned and support insightful explanations.

Activation-level uncertainty in deep neural networks

- Pablo Morales-Alvarez, Daniel Hernández-Lobato, Rafael Molina, José Miguel Hernández-Lobato
- abstract@open-review(Poster): Current approaches for uncertainty estimation in deep learning often produce too confident results. Bayesian Neural Networks (BNNs) model uncertainty in the space of weights, which is usually high-dimensional and limits the quality of variational approximations. The more recent functional BNNs (fBNNs) address this only partially because, although the prior is specified in the space of functions, the posterior approximation is still defined in terms of stochastic weights. In this work we propose to move uncertainty from the weights (which are deterministic) to the activation function. Specifically, the activations are modelled with simple 1D Gaussian Processes (GP), for which a triangular kernel inspired by the ReLu non-linearity is explored. Our experiments show that activation-level stochasticity provides more reliable uncertainty estimates than BNN and fBNN, whereas it performs competitively in standard prediction tasks. We also study the connection with deep GPs, both theoretically and empirically. More precisely, we show that activation-level uncertainty requires fewer inducing points and is better suited for deep architectures.

SkipW: Resource Adaptable RNN with Strict Upper Computational Limit

- Tsiry Mayet, Anne Lambert, Pascal Leguyadec, Francoise Le Bolzer, François Schnitzler
- abstract@open-review(Poster): We introduce Skip-Window, a method to allow recurrent neural networks (RNNs) to trade off accuracy for computational cost during the analysis of a sequence. Similarly to existing approaches, Skip-Window extends existing RNN cells by adding a mechanism to encourage the model to process fewer inputs. Unlike existing approaches, Skip-Window is able to respect a strict computational budget, making this model more suitable for limited hardware. We evaluate this approach on two datasets: a human activity recognition task and adding task. Our results show that Skip-Window is able to exceed the accuracy of existing approaches for a lower computational cost while strictly limiting said cost.

Wasserstein-2 Generative Networks

- Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, Evgeny Burnaev
- abstract@open-review(Poster): We propose a novel end-to-end non-minimax algorithm for training optimal transport mappings for the quadratic cost (Wasserstein-2 distance). The algorithm uses input convex neural networks and a cycle-consistency regularization to approximate Wasserstein-2 distance.

In contrast to popular entropic and quadratic regularizers, cycle-consistency does not introduce bias and scales well to high dimensions. From the theoretical side, we estimate the properties of the generative mapping fitted by our algorithm. From the practical side, we evaluate our algorithm on a wide range of tasks: image-to-image color transfer, latent space optimal transport, image-to-image style transfer, and domain adaptation.

Group Equivariant Stand-Alone Self-Attention For Vision

- David W. Romero, Jean-Baptiste Cordonnier
- abstract@open-review(Poster): We provide a general self-attention formulation to impose group equivariance to arbitrary symmetry groups. This is achieved by defining positional encodings that are invariant to the action of the group considered. Since the group acts on the positional encoding directly, group equivariant self-attention networks (GSA-Nets) are steerable by nature. Our experiments on vision benchmarks demonstrate consistent improvements of GSA-Nets over non-equivariant self-attention networks.

Continuous Wasserstein-2 Barycenter Estimation without Minimax Optimization

- Alexander Korotin, Lingxiao Li, Justin Solomon, Evgeny Burnaev
- abstract@open-review(Poster): Wasserstein barycenters provide a geometric notion of the weighted average of probability measures based on optimal transport. In this paper, we present a scalable algorithm to compute Wasserstein-2 barycenters given sample access to the input measures, which are not restricted to being discrete. While past approaches rely on entropic or quadratic regularization, we employ input convex neural networks and cycle-consistency regularization to avoid introducing bias. As a result, our approach does not resort to minimax optimization. We provide theoretical analysis on error bounds as well as empirical evidence of the effectiveness of the proposed approach in low-dimensional qualitative scenarios and high-dimensional quantitative experiments.

Identifying nonlinear dynamical systems with multiple time scales and long-range dependencies

- Dominik Schmidt, Georgia Koppe, Zahra Monfared, Max Beutelspacher, Daniel Durstewitz
- abstract@open-review(Spotlight): A main theoretical interest in biology and physics is to identify the nonlinear dynamical system (DS) that generated observed time series. Recurrent Neural Networks (RNN) are, in principle, powerful enough to approximate any underlying DS, but in their vanilla form suffer from the exploding vs. vanishing gradients problem. Previous attempts to alleviate this problem resulted either in more complicated, mathematically less tractable RNN architectures, or strongly limited the dynamical expressiveness of the RNN. Here we address this issue by suggesting a simple regularization scheme for vanilla RNN with ReLU activation which enables them to solve long-range dependency problems and express slow time scales, while retaining a simple mathematical structure which makes their DS properties partly analytically accessible. We prove two theorems that establish a tight connection between the regularized RNN dynamics and their gradients, illustrate on DS benchmarks that our regularization approach strongly eases the reconstruction of DS which harbor widely differing time scales, and show that our method is also en par with other long-range architectures like LSTMs on several tasks.

RNNLogic: Learning Logic Rules for Reasoning on Knowledge Graphs

- Meng Qu, Junkun Chen, Louis-Pascal Xhonneux, Yoshua Bengio, Jian Tang
- abstract@open-review(Poster): This paper studies learning logic rules for reasoning on knowledge graphs. Logic rules provide interpretable explanations when used for prediction as well as being able to generalize to other tasks, and hence are critical to learn. Existing methods either suffer from the problem of searching in a large search space (e.g., neural logic programming) or ineffective optimization due to sparse rewards (e.g., techniques based on reinforcement learning). To address these limitations, this paper proposes a probabilistic model called RNNLogic. RNNLogic treats logic rules as a latent variable, and simultaneously trains a rule generator as well as a reasoning predictor with logic rules. We develop an EM-based algorithm for optimization. In each iteration, the reasoning predictor is updated to explore some generated logic rules for reasoning. Then in the E-step, we select a set of high-quality rules from all generated rules with both the rule generator and reasoning predictor via posterior inference; and in the M-step, the rule generator is updated with the rules selected in the E-step. Experiments on four datasets prove the effectiveness of RNNLogic.

Selective Classification Can Magnify Disparities Across Groups

- Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, Percy Liang
- abstract@open-review(Poster): Selective classification, in which models can abstain on uncertain predictions, is a natural approach to improving accuracy in settings where errors are costly but abstentions are manageable. In this paper, we find that while selective classification can improve average accuracies, it can simultaneously magnify existing accuracy disparities between various groups within a population, especially in the presence of spurious correlations. We observe this behavior consistently across five vision and NLP datasets. Surprisingly, increasing abstentions can even decrease accuracies on some groups. To better understand this phenomenon, we study the margin distribution, which captures the model's confidences over all predictions. For symmetric margin distributions, we prove that whether selective classification monotonically improves or worsens accuracy is fully determined by the accuracy at full coverage (i.e., without any abstentions) and whether the distribution satisfies a property we call left-log-concavity. Our analysis also shows that selective classification tends to magnify full-coverage accuracy disparities. Motivated by our analysis, we train distributionally-robust models that achieve similar full-coverage accuracies across groups and show that selective classification uniformly improves each group on these models. Altogether, our results suggest that selective classification should be used with care and underscore the importance of training models to perform equally well across groups at full coverage.

FedMix: Approximation of Mixup under Mean Augmented Federated Learning

- Tehrim Yoon, Sumin Shin, Sung Ju Hwang, Eunho Yang
- abstract@open-review(Poster): Federated learning (FL) allows edge devices to collectively learn a model without directly sharing data within each device, thus preserving privacy and eliminating the need to store data globally. While there are promising results under the assumption of independent and identically distributed (iid) local data, current state-of-the-art algorithms suffer a performance degradation as the heterogeneity of local data across clients increases. To resolve this issue, we propose a simple framework, \emph{Mean Augmented Federated Learning (MAFL)}, where clients send and receive \emph{averaged} local data, subject to the privacy requirements of target applications. Under our framework, we propose a new augmentation algorithm, named \emph{FedMix}, which is inspired by a phenomenal yet simple data augmentation method, Mixup, but does not require local raw data to be directly shared among devices. Our method shows greatly improved performance in the standard benchmark datasets of FL, under highly non-iid federated settings, compared to conventional algorithms.

In-N-Out: Pre-Training and Self-Training using Auxiliary Information for Out-of-Distribution Robustness

- Sang Michael Xie, Ananya Kumar, Robbie Jones, Fereshte Khani, Tengyu Ma, Percy Liang
- abstract@open-review(Poster): Consider a prediction setting with few in-distribution labeled examples and many unlabeled examples both in- and out-of-distribution (OOD). The goal is to learn a model which performs well both in-distribution and OOD. In these settings, auxiliary information is often cheaply available for every input. How should we best leverage this auxiliary information for the prediction task? Empirically across three image and time-series datasets, and theoretically in a multi-task linear regression setting, we show that (i) using auxiliary information as input features improves in-

distribution error but can hurt OOD error; but (ii) using auxiliary information as outputs of auxiliary pre-training tasks improves OOD error. To get the best of both worlds, we introduce In-N-Out, which first trains a model with auxiliary inputs and uses it to pseudolabel all the in-distribution inputs, then pre-trains a model on OOD auxiliary outputs and fine-tunes this model with the pseudolabels (self-training). We show both theoretically and empirically that In-N-Out outperforms auxiliary inputs or outputs alone on both in-distribution and OOD error.

Sample-Efficient Automated Deep Reinforcement Learning

- Jörg K.H. Franke, Gregor Koehler, André Biedenkapp, Frank Hutter
- abstract@open-review(Poster): Despite significant progress in challenging problems across various domains, applying state-of-the-art deep reinforcement learning (RL) algorithms remains challenging due to their sensitivity to the choice of hyperparameters. This sensitivity can partly be attributed to the non-stationarity of the RL problem, potentially requiring different hyperparameter settings at various stages of the learning process. Additionally, in the RL setting, hyperparameter optimization (HPO) requires a large number of environment interactions, hindering the transfer of the successes in RL to real-world applications. In this work, we tackle the issues of sample-efficient and dynamic HPO in RL. We propose a population-based automated RL (AutoRL) framework to meta-optimize arbitrary off-policy RL algorithms. In this framework, we optimize the hyperparameters and also the neural architecture while simultaneously training the agent. By sharing the collected experience across the population, we substantially increase the sample efficiency of the meta-optimization. We demonstrate the capabilities of our sample-efficient AutoRL approach in a case study with the popular TD3 algorithm in the MuJoCo benchmark suite, where we reduce the number of environment interactions needed for meta-optimization by up to an order of magnitude compared to population-based training.

A Temporal Kernel Approach for Deep Learning with Continuous-time Information

- Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, Kannan Achan
- abstract@open-review(Poster): Sequential deep learning models such as RNN, causal CNN and attention mechanism do not readily consume continuous-time information. Discretizing the temporal data, as we show, causes inconsistency even for simple continuous-time processes. Current approaches often handle time in a heuristic manner to be consistent with the existing deep learning architectures and implementations. In this paper, we provide a principled way to characterize continuous-time systems using deep learning tools. Notably, the proposed approach applies to all the major deep learning architectures and requires little modifications to the implementation. The critical insight is to represent the continuous-time system by composing neural networks with a temporal kernel, where we gain our intuition from the recent advancements in understanding deep learning with Gaussian process and neural tangent kernel. To represent the temporal kernel, we introduce the random feature approach and convert the kernel learning problem to spectral density estimation under reparameterization. We further prove the convergence and consistency results even when the temporal kernel is non-stationary, and the spectral density is misspecified. The simulations and real-data experiments demonstrate the empirical effectiveness of our temporal kernel approach in a broad range of settings.

Convex Regularization behind Neural Reconstruction

- Arda Sahiner, Morteza Mardani, Batu Ozturkler, Mert Pilanci, John M. Pauly
- abstract@open-review(Poster): Neural networks have shown tremendous potential for reconstructing high-resolution images in inverse problems. The non-convex and opaque nature of neural networks, however, hinders their utility in sensitive applications such as medical imaging. To cope with this challenge, this paper advocates a convex duality framework that makes a two-layer fully-convolutional ReLU denoising network amenable to convex optimization. The convex dual network not only offers the optimum training with convex solvers, but also facilitates interpreting training and prediction. In particular, it implies training neural networks with weight decay regularization induces path sparsity while the prediction is piecewise linear filtering. A range of experiments with MNIST and fastMRI datasets confirm the efficacy of the dual network optimization problem.

Vector-output ReLU Neural Network Problems are Copositive Programs: Convex Analysis of Two Layer Networks and Polynomial-time Algorithms

- Arda Sahiner, Tolga Ergen, John M. Pauly, Mert Pilanci
- abstract@open-review(Poster): We describe the convex semi-infinite dual of the two-layer vector-output ReLU neural network training problem. This semi-infinite dual admits a finite dimensional representation, but its support is over a convex set which is difficult to characterize. In particular, we demonstrate that the non-convex neural network training problem is equivalent to a finite-dimensional convex copositive program. Our work is the first to identify this strong connection between the global optima of neural networks and those of copositive programs. We thus demonstrate how neural networks implicitly attempt to solve copositive programs via semi-nonnegative matrix factorization, and draw key insights from this formulation. We describe the first algorithms for provably finding the global minimum of the vector output neural network training problem, which are polynomial in the number of samples for a fixed data rank, yet exponential in the dimension. However, in the case of convolutional architectures, the computational complexity is exponential in only the filter size and polynomial in all other parameters. We describe the circumstances in which we can find the global optimum of this neural network training problem exactly with soft-thresholded SVD, and provide a copositive relaxation which is guaranteed to be exact for certain classes of problems, and which corresponds with the solution of Stochastic Gradient Descent in practice.

Learning Better Structured Representations Using Low-rank Adaptive Label Smoothing

- Asish Ghoshal, Xilun Chen, Sonal Gupta, Luke Zettlemoyer, Yashar Mehdad
- abstract@open-review(Poster): Training with soft targets instead of hard targets has been shown to improve performance and calibration of deep neural networks. Label smoothing is a popular way of computing soft targets, where one-hot encoding of a class is smoothed with a uniform distribution. Owing to its simplicity, label smoothing has found wide-spread use for training deep neural networks on a wide variety of tasks, ranging from image and text classification to machine translation and semantic parsing. Complementing recent empirical justification for label smoothing, we obtain PAC-Bayesian generalization bounds for label smoothing and show that the generalization error depends on the choice of the noise (smoothing) distribution. Then we propose low-rank adaptive label smoothing (LORAS): a simple yet novel method for training with learned soft targets that generalizes label smoothing and adapts to the latent structure of the label space in structured prediction tasks. Specifically, we evaluate our method on semantic parsing tasks and show that training with appropriately smoothed soft targets can significantly improve accuracy and model calibration, especially in low-resource settings. Used in conjunction with pre-trained sequence-to-sequence models, our method achieves state of the art performance on four semantic parsing data sets. LORAS can be used with any model, improves performance and implicit model calibration without increasing the number of model parameters, and can be scaled to problems with large label spaces containing tens of thousands of labels.

Understanding the role of importance weighting for deep learning

- Da Xu, Yuting Ye, Chuanwei Ruan
- abstract@open-review(Spotlight): The recent paper by Byrd & Lipton (2019), based on empirical observations, raises a major concern on the impact of importance weighting for the over-parameterized deep learning models. They observe that as long as the model can separate the training data, the impact of importance weighting diminishes as the training proceeds. Nevertheless, there lacks a rigorous characterization of this phenomenon. In this paper, we provide formal characterizations and theoretical justifications on the role of importance weighting with respect to the implicit bias of gradient descent and margin-based learning theory. We reveal both the optimization dynamics and generalization performance under deep learning models. Our work not only

explains the various novel phenomenons observed for importance weighting in deep learning, but also extends to the studies where the weights are being optimized as part of the model, which applies to a number of topics under active research.

Training GANs with Stronger Augmentations via Contrastive Discriminator

- Jongheon Jeong, Jinwoo Shin
- abstract@open-review(Poster): Recent works in Generative Adversarial Networks (GANs) are actively revisiting various data augmentation techniques as an effective way to prevent discriminator overfitting. It is still unclear, however, that which augmentations could actually improve GANs, and in particular, how to apply a wider range of augmentations in training. In this paper, we propose a novel way to address these questions by incorporating a recent contrastive representation learning scheme into the GAN discriminator, coined ContraD. This "fusion" enables the discriminators to work with much stronger augmentations without increasing their training instability, thereby preventing the discriminator overfitting issue in GANs more effectively. Even better, we observe that the contrastive learning itself also benefits from our GAN training, i.e., by maintaining discriminative features between real and fake samples, suggesting a strong coherence between the two worlds: good contrastive representations are also good for GAN discriminators, and vice versa. Our experimental results show that GANs with ContraD consistently improve FID and IS compared to other recent techniques incorporating data augmentations, still maintaining highly discriminative features in the discriminator in terms of the linear evaluation. Finally, as a byproduct, we also show that our GANs trained in an unsupervised manner (without labels) can induce many conditional generative models via a simple latent sampling, leveraging the learned features of ContraD. Code is available at https://github.com/jh-jeong/ContraD.

Private Image Reconstruction from System Side Channels Using Generative Models

- Yuanyuan Yuan, Shuai Wang, Junping Zhang
- abstract@open-review(Poster): System side channels denote effects imposed on the underlying system and hardware when running a program, such as its accessed CPU cache lines. Side channel analysis (SCA) allows attackers to infer program secrets based on observed side channel signals. Given the ever-growing adoption of machine learning as a service (MLaaS), image analysis software on cloud platforms has been exploited by reconstructing private user images from system side channels. Nevertheless, to date, SCA is still highly challenging, requiring technical knowledge of victim software's internal operations. For existing SCA attacks, comprehending such internal operations requires heavyweight program analysis or manual efforts.

This research proposes an attack framework to reconstruct private user images processed by media software via system side channels. The framework forms an effective workflow by incorporating convolutional networks, variational autoencoders, and generative adversarial networks. Our evaluation of two popular side channels shows that the reconstructed images consistently match user inputs, making privacy leakage attacks more practical. We also show surprising results that even one-bit data read/write pattern side channels, which are deemed minimally informative, can be used to reconstruct quality images using our framework.

RMSprop converges with proper hyper-parameter

- Naichen Shi, Dawei Li, Mingyi Hong, Ruoyu Sun
- abstract@open-review(Spotlight): Despite the existence of divergence examples, RMSprop remains one of the most popular algorithms in machine learning. Towards closing the gap between theory and practice, we prove that RMSprop converges with proper choice of hyper-parameters under certain conditions. More specifically, we prove that when the hyper-parameter β_2 is close enough to 1, RMSprop and its random shuffling version converge to a bounded region in general, and to critical points in the interpolation regime. It is worth mentioning that our results do not depend on ``bounded gradient" assumption, which is often the key assumption utilized by existing theoretical work for Adam-type adaptive gradient method. Removing this assumption allows us to establish a phase transition from divergence to non-divergence for RMSprop.

Finally, based on our theory, we conjecture that in practice there is a critical threshold $\sf{\beta_2^*}$, *such that RMSprop generates reasonably good results only if $1>\beta_2\ge \sf{\beta_2^*}$*. We provide empirical evidence for such a phase transition in our numerical experiments.

Learning to Make Decisions via Submodular Regularization

- Ayya Alieva, Aiden Aceves, Jialin Song, Stephen Mayo, Yisong Yue, Yuxin Chen
- abstract@open-review(Poster): Many sequential decision making tasks can be viewed as combinatorial optimization problems over a large number of actions. When the cost of evaluating an action is high, even a greedy algorithm, which iteratively picks the best action given the history, is prohibitive to run. In this paper, we aim to learn a greedy heuristic for sequentially selecting actions as a surrogate for invoking the expensive oracle when evaluating an action. In particular, we focus on a class of combinatorial problems that can be solved via submodular maximization (either directly on the objective function or via submodular surrogates). We introduce a data-driven optimization framework based on the submodular-norm loss, a novel loss function that encourages the resulting objective to exhibit diminishing returns. Our framework outputs a surrogate objective that is efficient to train, approximately submodular, and can be made permutation-invariant. The latter two properties allow us to prove strong approximation guarantees for the learned greedy heuristic. Furthermore, we show that our model can be easily integrated with modern deep imitation learning pipelines for sequential prediction tasks. We demonstrate the performance of our algorithm on a variety of batched and sequential optimization tasks, including set cover, active learning, and Bayesian optimization for protein engineering.

The Recurrent Neural Tangent Kernel

- Sina Alemohammad, Zichao Wang, Randall Balestriero, Richard Baraniuk
- abstract@open-review(Poster): The study of deep neural networks (DNNs) in the infinite-width limit, via the so-called neural tangent kernel (NTK) approach, has provided new insights into the dynamics of learning, generalization, and the impact of initialization. One key DNN architecture remains to be kernelized, namely, the recurrent neural network (RNN). In this paper we introduce and study the Recurrent Neural Tangent Kernel (RNTK), which provides new insights into the behavior of overparametrized RNNs. A key property of the RNTK should greatly benefit practitioners is its ability to compare inputs of different length. To this end, we characterize how the RNTK weights different time steps to form its output under different initialization parameters and nonlinearity choices. A synthetic and 56 real-world data experiments demonstrate that the RNTK offers significant performance gains over other kernels, including standard NTKs, across a wide array of data sets.

Why Are Convolutional Nets More Sample-Efficient than Fully-Connected Nets?

- Zhiyuan Li, Yi Zhang, Sanjeev Arora
- abstract@open-review(Oral): Convolutional neural networks often dominate fully-connected counterparts in generalization performance, especially on image classification tasks. This is often explained in terms of \textquotedblleft better inductive bias.\textquotedblright\ However, this has not been made mathematically rigorous, and the hurdle is that the sufficiently wide fully-connected net can always simulate the convolutional net. Thus the training algorithm plays a role. The current work describes a natural task on which a provable sample complexity gap can be shown, for standard training algorithms. We construct a single natural distribution on $\mathbb{R}^d\times{\pm 1}$ on which any orthogonal-invariant algorithm (i.e. fully-connected networks trained with most gradient-based methods from gaussian initialization) requires $\Omega(d^2)$ samples to generalize while $O(1)$ samples suffice for convolutional architectures. Furthermore, we demonstrate a single target function, learning which on all possible distributions leads to an $O(1)$ vs $\Omega(d^2/\varepsilon)$ gap. The proof relies on the fact that SGD on fully-connected network is orthogonal equivariant. Similar results are achieved for ℓ_2 regression and adaptive training algorithms, e.g. Adam and AdaGrad, which are only permutation equivariant.

Evaluation of Similarity-based Explanations

- Kazuaki Hanawa, Sho Yokoi, Satoshi Hara, Kentaro Inui
- abstract@open-review(Poster): Explaining the predictions made by complex machine learning models helps users to understand and accept the predicted outputs with confidence. One promising way is to use similarity-based explanation that provides similar instances as evidence to support model predictions. Several relevance metrics are used for this purpose. In this study, we investigated relevance metrics that can provide reasonable explanations to users. Specifically, we adopted three tests to evaluate whether the relevance metrics satisfy the minimal requirements for similarity-based explanation. Our experiments revealed that the cosine similarity of the gradients of the loss performs best, which would be a recommended choice in practice. In addition, we showed that some metrics perform poorly in our tests and analyzed the reasons of their failure. We expect our insights to help practitioners in selecting appropriate relevance metrics and also aid further researches for designing better relevance metrics for explanations.

Adaptive Procedural Task Generation for Hard-Exploration Problems

- Kuan Fang, Yuke Zhu, Silvio Savarese, L. Fei-Fei
- abstract@open-review(Poster): We introduce Adaptive Procedural Task Generation (APT-Gen), an approach to progressively generate a sequence of tasks as curricula to facilitate reinforcement learning in hard-exploration problems. At the heart of our approach, a task generator learns to create tasks from a parameterized task space via a black-box procedural generation module. To enable curriculum learning in the absence of a direct indicator of learning progress, we propose to train the task generator by balancing the agent's performance in the generated tasks and the similarity to the target tasks. Through adversarial training, the task similarity is adaptively estimated by a task discriminator defined on the agent's experiences, allowing the generated tasks to approximate target tasks of unknown parameterization or outside of the predefined task space. Our experiments on the grid world and robotic manipulation task domains show that APT-Gen achieves substantially better performance than various existing baselines by generating suitable tasks of rich variations.

Linear Last-iterate Convergence in Constrained Saddle-point Optimization

- Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, Haipeng Luo
- abstract@open-review(Poster): Optimistic Gradient Descent Ascent (OGDA) and Optimistic Multiplicative Weights Update (OMWU) for saddle-point optimization have received growing attention due to their favorable last-iterate convergence. However, their behaviors for simple bilinear games over the probability simplex are still not fully understood --- previous analysis lacks explicit convergence rates, only applies to an exponentially small learning rate, or requires additional assumptions such as the uniqueness of the optimal solution.

In this work, we significantly expand the understanding of last-iterate convergence for OGDA and OMWU in the constrained setting. Specifically, for OMWU in bilinear games over the simplex, we show that when the equilibrium is unique, linear last-iterate convergence is achievable with a constant learning rate, which improves the result of (Daskalakis & Panageas, 2019) under the same assumption. We then significantly extend the results to more general objectives and feasible sets for the projected OGDA algorithm, by introducing a sufficient condition under which OGDA exhibits concrete last-iterate convergence rates with a constant learning rate. We show that bilinear games over any polytope satisfy this condition and OGDA converges exponentially fast even without the unique equilibrium assumption. Our condition also holds for strongly-convex-strongly-concave functions, recovering the result of (Hsieh et al., 2019). Finally, we provide experimental results to further support our theory.

On Graph Neural Networks versus Graph-Augmented MLPs

- Lei Chen, Zhengdao Chen, Joan Bruna
- abstract@open-review(Poster): From the perspectives of expressive power and learning, this work compares multi-layer Graph Neural Networks (GNNs) with a simplified alternative that we call Graph-Augmented Multi-Layer Perceptrons (GA-MLPs), which first augments node features with certain multi-hop operators on the graph and then applies learnable node-wise functions. From the perspective of graph isomorphism testing, we show both theoretically and numerically that GA-MLPs with suitable operators can distinguish almost all non-isomorphic graphs, just like the Weisfeiler-Lehman (WL) test and GNNs. However, by viewing them as node-level functions and examining the equivalence classes they induce on rooted graphs, we prove a separation in expressive power between GA-MLPs and GNNs that grows exponentially in depth. In particular, unlike GNNs, GA-MLPs are unable to count the number of attributed walks. We also demonstrate via community detection experiments that GA-MLPs can be limited by their choice of operator family, whereas GNNs have higher flexibility in learning.

Solving Compositional Reinforcement Learning Problems via Task Reduction

- Yunfei Li, Yilin Wu, Huazhe Xu, Xiaolong Wang, Yi Wu
- abstract@open-review(Poster): We propose a novel learning paradigm, Self-Imitation via Reduction (SIR), for solving compositional reinforcement learning problems. SIR is based on two core ideas: task reduction and self-imitation. Task reduction tackles a hard-to-solve task by actively reducing it to an easier task whose solution is known by the RL agent. Once the original hard task is successfully solved by task reduction, the agent naturally obtains a self-generated solution trajectory to imitate. By continuously collecting and imitating such demonstrations, the agent is able to progressively expand the solved subspace in the entire task space. Experiment results show that SIR can significantly accelerate and improve learning on a variety of challenging sparse-reward continuous-control problems with compositional structures. Code and videos are available at https://sites.google.com/view/sir-compositional.

The Intrinsic Dimension of Images and Its Impact on Learning

- Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, Tom Goldstein
- abstract@open-review(Spotlight): It is widely believed that natural image data exhibits low-dimensional structure despite the high dimensionality of conventional pixel representations. This idea underlies a common intuition for the remarkable success of deep learning in computer vision. In this work, we apply dimension estimation tools to popular datasets and investigate the role of low-dimensional structure in deep learning. We find that common natural image datasets indeed have very low intrinsic dimension relative to the high number of pixels in the images. Additionally, we find that low dimensional datasets are easier for neural networks to learn, and models solving these tasks generalize better from training to test data. Along the way, we develop a technique for validating our dimension estimation tools on synthetic data generated by GANs allowing us to actively manipulate the intrinsic dimension by controlling the image generation process. Code for our experiments may be found \href{https://github.com/ppope/dimensions}{here}.

Conditional Generative Modeling via Learning the Latent Space

- Sameera Ramasinghe, Kanchana Nisal Ranasinghe, Salman Khan, Nick Barnes, Stephen Gould
- abstract@open-review(Poster): Although deep learning has achieved appealing results on several machine learning tasks, most of the models are deterministic at inference, limiting their application to single-modal settings. We propose a novel general-purpose framework for conditional generation in multimodal spaces, that uses latent variables to model generalizable learning patterns while minimizing a family of regression cost functions. At inference, the latent variables are optimized to find solutions corresponding to multiple output modes. Compared to existing generative solutions, our approach demonstrates faster and more stable convergence, and can learn better representations for downstream tasks. Importantly, it provides a simple generic

model that can perform better than highly engineered pipelines tailored using domain expertise on a variety of tasks, while generating diverse outputs. Code available at https://github.com/samgregoost/cGML.

Learning with Feature-Dependent Label Noise: A Progressive Approach

- Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, Chao Chen
- abstract@open-review(Spotlight): Label noise is frequently observed in real-world large-scale datasets. The noise is introduced due to a variety of reasons; it is heterogeneous and feature-dependent. Most existing approaches to handling noisy labels fall into two categories: they either assume an ideal feature-independent noise, or remain heuristic without theoretical guarantees. In this paper, we propose to target a new family of feature-dependent label noise, which is much more general than commonly used i.i.d. label noise and encompasses a broad spectrum of noise patterns. Focusing on this general noise family, we propose a progressive label correction algorithm that iteratively corrects labels and refines the model. We provide theoretical guarantees showing that for a wide variety of (unknown) noise patterns, a classifier trained with this strategy converges to be consistent with the Bayes classifier. In experiments, our method outperforms SOTA baselines and is robust to various noise types and levels.

DialoGraph: Incorporating Interpretable Strategy-Graph Networks into Negotiation Dialogues

- Rishabh Joshi, Vidhisha Balachandran, Shikhar Vashishth, Alan Black, Yulia Tsvetkov
- abstract@open-review(Poster): To successfully negotiate a deal, it is not enough to communicate fluently: pragmatic planning of persuasive negotiation strategies is essential. While modern dialogue agents excel at generating fluent sentences, they still lack pragmatic grounding and cannot reason strategically. We present DialoGraph, a negotiation system that incorporates pragmatic strategies in a negotiation dialogue using graph neural networks. DialoGraph explicitly incorporates dependencies between sequences of strategies to enable improved and interpretable prediction of next optimal strategies, given the dialogue context. Our graph-based method outperforms prior state-of-the-art negotiation models both in the accuracy of strategy/dialogue act prediction and in the quality of downstream dialogue response generation. We qualitatively show further benefits of learned strategy-graphs in providing explicit associations between effective negotiation strategies over the course of the dialogue, leading to interpretable and strategic dialogues.

WaNet - Imperceptible Warping-based Backdoor Attack

- Tuan Anh Nguyen, Anh Tuan Tran
- abstract@open-review(Poster): With the thriving of deep learning and the widespread practice of using pre-trained networks, backdoor attacks have become an increasing security threat drawing many research interests in recent years. A third-party model can be poisoned in training to work well in normal conditions but behave maliciously when a trigger pattern appears. However, the existing backdoor attacks are all built on noise perturbation triggers, making them noticeable to humans. In this paper, we instead propose using warping-based triggers. The proposed backdoor outperforms the previous methods in a human inspection test by a wide margin, proving its stealthiness. To make such models undetectable by machine defenders, we propose a novel training mode, called the ``noise mode. The trained networks successfully attack and bypass the state-ofthe art defense methods on standard classification datasets, including MNIST, CIFAR-10, GTSRB, and CelebA. Behavior analyses show that our backdoors are transparent to network inspection, further proving this novel attack mechanism's efficiency.

Nonseparable Symplectic Neural Networks

- Shiying Xiong, Yunjin Tong, Xingzhe He, Shuqi Yang, Cheng Yang, Bo Zhu
- abstract@open-review(Poster): Predicting the behaviors of Hamiltonian systems has been drawing increasing attention in scientific machine learning. However, the vast majority of the literature was focused on predicting separable Hamiltonian systems with their kinematic and potential energy terms being explicitly decoupled, while building data-driven paradigms to predict nonseparable Hamiltonian systems that are ubiquitous in fluid dynamics and quantum mechanics were rarely explored. The main computational challenge lies in the effective embedding of symplectic priors to describe the inherently coupled evolution of position and momentum, which typically exhibits intricate dynamics. To solve the problem, we propose a novel neural network architecture, Nonseparable Symplectic Neural Networks (NSSNNs), to uncover and embed the symplectic structure of a nonseparable Hamiltonian system from limited observation data. The enabling mechanics of our approach is an augmented symplectic time integrator to decouple the position and momentum energy terms and facilitate their evolution. We demonstrated the efficacy and versatility of our method by predicting a wide range of Hamiltonian systems, both separable and nonseparable, including chaotic vortical flows. We showed the unique computational merits of our approach to yield long-term, accurate, and robust predictions for large-scale Hamiltonian systems by rigorously enforcing symplectomorphism.

Discovering Diverse Multi-Agent Strategic Behavior via Reward Randomization

- Zhenggang Tang, Chao Yu, Boyuan Chen, Huazhe Xu, Xiaolong Wang, Fei Fang, Simon Shaolei Du, Yu Wang, Yi Wu
- abstract@open-review(Poster): We propose a simple, general and effective technique, Reward Randomization for discovering diverse strategic policies in complex multi-agent games. Combining reward randomization and policy gradient, we derive a new algorithm, Reward-Randomized Policy Gradient (RPG). RPG is able to discover a set of multiple distinctive human-interpretable strategies in challenging temporal trust dilemmas, including grid-world games and a real-world game Agar.io, where multiple equilibria exist but standard multi-agent policy gradient algorithms always converge to a fixed one with a sub-optimal payoff for every player even using state-of-the-art exploration techniques. Furthermore, with the set of diverse strategies from RPG, we can (1) achieve higher payoffs by fine-tuning the best policy from the set; and (2) obtain an adaptive agent by using this set of strategies as its training opponents.

Multi-timescale Representation Learning in LSTM Language Models

- Shivangi Mahto, Vy Ai Vo, Javier S. Turek, Alexander Huth
- abstract@open-review(Poster): Language models must capture statistical dependencies between words at timescales ranging from very short to very long. Earlier work has demonstrated that dependencies in natural language tend to decay with distance between words according to a power law. However, it is unclear how this knowledge can be used for analyzing or designing neural network language models. In this work, we derived a theory for how the memory gating mechanism in long short-term memory (LSTM) language models can capture power law decay. We found that unit timescales within an LSTM, which are determined by the forget gate bias, should follow an Inverse Gamma distribution. Experiments then showed that LSTM language models trained on natural English text learn to approximate this theoretical distribution. Further, we found that explicitly imposing the theoretical distribution upon the model during training yielded better language model perplexity overall, with particular improvements for predicting low-frequency (rare) words. Moreover, the explicit multi-timescale model selectively routes information about different types of words through units with different timescales, potentially improving model interpretability. These results demonstrate the importance of careful, theoretically-motivated analysis of memory and timescale in language models.

Explaining the Efficacy of Counterfactually Augmented Data

- Divyansh Kaushik, Amrith Setlur, Eduard H Hovy, Zachary Chase Lipton

- abstract@open-review(Poster): In attempts to produce machine learning models less reliant on spurious patterns in NLP datasets, researchers have recently proposed curating counterfactually augmented data (CAD) via a human-in-the-loop process in which given some documents and their (initial) labels, humans must revise the text to make a counterfactual label applicable. Importantly, edits that are not necessary to flip the applicable label are prohibited. Models trained on the augmented (original and revised) data appear, empirically, to rely less on semantically irrelevant words and to generalize better out of domain. While this work draws loosely on causal thinking, the underlying causal model (even at an abstract level) and the principles underlying the observed out-of-domain improvements remain unclear. In this paper, we introduce a toy analog based on linear Gaussian models, observing interesting relationships between causal models, measurement noise, out-of-domain generalization, and reliance on spurious signals. Our analysis provides some insights that help to explain the efficacy of CAD. Moreover, we develop the hypothesis that while adding noise to causal features should degrade both in-domain and out-of-domain performance, adding noise to non-causal features should lead to relative improvements in out-of-domain performance. This idea inspires a speculative test for determining whether a feature attribution technique has identified the causal spans. If adding noise (e.g., by random word flips) to the highlighted spans degrades both in-domain and out-of-domain performance on a battery of challenge datasets, but adding noise to the complement gives improvements out-of-domain, this suggests we have identified causal spans. Thus, we present a large scale empirical study comparing spans edited to create CAD to those selected by attention and saliency maps. Across numerous challenge domains and models, we find that the hypothesized phenomenon is pronounced for CAD.

Revisiting Locally Supervised Learning: an Alternative to End-to-end Training

- Yulin Wang, Zanlin Ni, Shiji Song, Le Yang, Gao Huang
- abstract@open-review(Poster): Due to the need to store the intermediate activations for back-propagation, end-to-end (E2E) training of deep networks usually suffers from high GPUs memory footprint. This paper aims to address this problem by revisiting the locally supervised learning, where a network is split into gradient-isolated modules and trained with local supervision. We experimentally show that simply training local modules with E2E loss tends to collapse task-relevant information at early layers, and hence hurts the performance of the full model. To avoid this issue, we propose an information propagation (InfoPro) loss, which encourages local modules to preserve as much useful information as possible, while progressively discard task-irrelevant information. As InfoPro loss is difficult to compute in its original form, we derive a feasible upper bound as a surrogate optimization objective, yielding a simple but effective algorithm. In fact, we show that the proposed method boils down to minimizing the combination of a reconstruction loss and a normal cross-entropy/contrastive term. Extensive empirical results on five datasets (i.e., CIFAR, SVHN, STL-10, ImageNet and Cityscapes) validate that InfoPro is capable of achieving competitive performance with less than 40% memory footprint compared to E2E training, while allowing using training data with higher-resolution or larger batch sizes under the same GPU memory constraint. Our method also enables training local modules asynchronously for potential training acceleration.

How Much Over-parameterization Is Sufficient to Learn Deep ReLU Networks?

- Zixiang Chen, Yuan Cao, Difan Zou, Quanquan Gu
- abstract@open-review(Poster): A recent line of research on deep learning focuses on the extremely over-parameterized setting, and shows that when the network width is larger than a high degree polynomial of the training sample size n and the inverse of the target error ϵ^{-1}, deep neural networks learned by (stochastic) gradient descent enjoy nice optimization and generalization guarantees. Very recently, it is shown that under certain margin assumptions on the training data, a polylogarithmic width condition suffices for two-layer ReLU networks to converge and generalize (Ji and Telgarsky, 2020). However, whether deep neural networks can be learned with such a mild over-parameterization is still an open question. In this work, we answer this question affirmatively and establish sharper learning guarantees for deep ReLU networks trained by (stochastic) gradient descent. In specific, under certain assumptions made in previous work, our optimization and generalization guarantees hold with network width polylogarithmic in n and ϵ^{-1}. Our results push the study of over-parameterized deep neural networks towards more practical settings.

Blending MPC & Value Function Approximation for Efficient Reinforcement Learning

- Mohak Bhardwaj, Sanjiban Choudhury, Byron Boots
- abstract@open-review(Poster): Model-Predictive Control (MPC) is a powerful tool for controlling complex, real-world systems that uses a model to make predictions about future behavior. For each state encountered, MPC solves an online optimization problem to choose a control action that will minimize future cost. This is a surprisingly effective strategy, but real-time performance requirements warrant the use of simple models. If the model is not sufficiently accurate, then the resulting controller can be biased, limiting performance. We present a framework for improving on MPC with model-free reinforcement learning (RL). The key insight is to view MPC as constructing a series of local Q-function approximations. We show that by using a parameter λ, similar to the trace decay parameter in TD(λ), we can systematically trade-off learned value estimates against the local Q-function approximations. We present a theoretical analysis that shows how error from inaccurate models in MPC and value function estimation in RL can be balanced. We further propose an algorithm that changes λ over time to reduce the dependence on MPC as our estimates of the value function improve, and test the efficacy our approach on challenging high-dimensional manipulation tasks with biased models in simulation. We demonstrate that our approach can obtain performance comparable with MPC with access to true dynamics even under severe model bias and is more sample efficient as compared to model-free RL.

Probabilistic Numeric Convolutional Neural Networks

- Marc Anton Finzi, Roberto Bondesan, Max Welling
- abstract@open-review(Poster): Continuous input signals like images and time series that are irregularly sampled or have missing values are challenging for existing deep learning methods. Coherently defined feature representations must depend on the values in unobserved regions of the input. Drawing from the work in probabilistic numerics, we propose Probabilistic Numeric Convolutional Neural Networks which represent features as Gaussian processes, providing a probabilistic description of discretization error. We then define a convolutional layer as the evolution of a PDE defined on this GP, followed by a nonlinearity. This approach also naturally admits steerable equivariant convolutions under e.g. the rotation group. In experiments we show that our approach yields a $3\times$ reduction of error from the previous state of the art on the SuperPixel-MNIST dataset and competitive performance on the medical time series dataset PhysioNet2012.

HeteroFL: Computation and Communication Efficient Federated Learning for Heterogeneous Clients

- Enmao Diao, Jie Ding, Vahid Tarokh
- abstract@open-review(Poster): Federated Learning (FL) is a method of training machine learning models on private data distributed over a large number of possibly heterogeneous clients such as mobile phones and IoT devices. In this work, we propose a new federated learning framework named HeteroFL to address heterogeneous clients equipped with very different computation and communication capabilities. Our solution can enable the training of heterogeneous local models with varying computation complexities and still produce a single global inference model. For the first time, our method challenges the underlying assumption of existing work that local models have to share the same architecture as the global model. We demonstrate several strategies to enhance FL training and conduct extensive empirical evaluations, including five computation complexity levels of three model architecture on three datasets. We show that adaptively distributing subnetworks according to clients' capabilities is both computation and communication efficient.

Semantic Re-tuning with Contrastive Tension

- Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, Magnus Sahlgren

- abstract@open-review(Poster): Extracting semantically useful natural language sentence representations from pre-trained deep neural networks such as Transformers remains a challenge. We first demonstrate that pre-training objectives impose a significant task bias onto the final layers of models with a layer-wise survey of the Semantic Textual Similarity (STS) correlations for multiple common Transformer language models. We then propose a new self-supervised method called Contrastive Tension (CT) to counter such biases. CT frames the training objective as a noise-contrastive task between the final layer representations of two independent models, in turn making the final layer representations suitable for feature extraction. Results from multiple common unsupervised and supervised STS tasks indicate that CT outperforms previous State Of The Art (SOTA), and when combining CT with supervised data we improve upon previous SOTA results with large margins.

Dataset Meta-Learning from Kernel Ridge-Regression

- Timothy Nguyen, Zhourong Chen, Jaehoon Lee
- abstract@open-review(Poster): One of the most fundamental aspects of any machine learning algorithm is the training data used by the algorithm. We introduce the novel concept of ϵ-approximation of datasets, obtaining datasets which are much smaller than or are significant corruptions of the original training data while maintaining similar performance. We introduce a meta-learning algorithm Kernel Inducing Points (KIP) for obtaining such remarkable datasets, drawing inspiration from recent developments in the correspondence between infinitely-wide neural networks and kernel ridge-regression (KRR). For KRR tasks, we demonstrate that KIP can compress datasets by one or two orders of magnitude, significantly improving previous dataset distillation and subset selection methods while obtaining state of the art results for MNIST and CIFAR10 classification. Furthermore, our KIP-learned datasets are transferable to the training of finite-width neural networks even beyond the lazy-training regime. Consequently, we obtain state of the art results for neural network dataset distillation with potential applications to privacy-preservation.

AUXILIARY TASK UPDATE DECOMPOSITION: THE GOOD, THE BAD AND THE NEUTRAL

- Lucio M. Dery, Yann Dauphin, David Grangier
- abstract@open-review(Poster): While deep learning has been very beneficial in data-rich settings, tasks with smaller training set often resort to pre-training or multitask learning to leverage data from other tasks. In this case, careful consideration is needed to select tasks and model parameterizations such that updates from the auxiliary tasks actually help the primary task. We seek to alleviate this burden by formulating a model-agnostic framework that performs fine-grained manipulation of the auxiliary task gradients. We propose to decompose auxiliary updates into directions which help, damage or leave the primary task loss unchanged. This allows weighting the update directions differently depending on their impact on the problem of interest. We present a novel and efficient algorithm for that purpose and show its advantage in practice. Our method leverages efficient automatic differentiation procedures and randomized singular value decomposition for scalability. We show that our framework is generic and encompasses some prior work as particular cases. Our approach consistently outperforms strong and widely used baselines when leveraging out-of-distribution data for Text and Image classification tasks.

Fast And Slow Learning Of Recurrent Independent Mechanisms

- Kanika Madan, Nan Rosemary Ke, Anirudh Goyal, Bernhard Schölkopf, Yoshua Bengio
- abstract@open-review(Poster): Decomposing knowledge into interchangeable pieces promises a generalization advantage when there are changes in distribution. A learning agent interacting with its environment is likely to be faced with situations requiring novel combinations of existing pieces of knowledge. We hypothesize that such a decomposition of knowledge is particularly relevant for being able to generalize in a systematic way to out-of-distribution changes. To study these ideas, we propose a particular training framework in which we assume that the pieces of knowledge an agent needs and its reward function are stationary and can be re-used across tasks. An attention mechanism dynamically selects which modules can be adapted to the current task, and the parameters of the \textit{selected} modules are allowed to change quickly as the learner is confronted with variations in what it experiences, while the parameters of the attention mechanisms act as stable, slowly changing, meta-parameters. We focus on pieces of knowledge captured by an ensemble of modules sparsely communicating with each other via a bottleneck of attention. We find that meta-learning the modular aspects of the proposed system greatly helps in achieving faster adaptation in a reinforcement learning setup involving navigation in a partially observed grid world with image-level input. We also find that reversing the role of parameters and meta-parameters does not work nearly as well, suggesting a particular role for fast adaptation of the dynamically selected modules.

Auction Learning as a Two-Player Game

- Jad Rahme, Samy Jelassi, S. Matthew Weinberg
- abstract@open-review(Poster): Designing an incentive compatible auction that maximizes expected revenue is a central problem in Auction Design. While theoretical approaches to the problem have hit some limits, a recent research direction initiated by Duetting et al. (2019) consists in building neural network architectures to find optimal auctions. We propose two conceptual deviations from their approach which result in enhanced performance. First, we use recent results in theoretical auction design to introduce a time-independent Lagrangian. This not only circumvents the need for an expensive hyper-parameter search (as in prior work), but also provides a single metric to compare the performance of two auctions (absent from prior work). Second, the optimization procedure in previous work uses an inner maximization loop to compute optimal misreports. We amortize this process through the introduction of an additional neural network. We demonstrate the effectiveness of our approach by learning competitive or strictly improved auctions compared to prior work. Both results together further imply a novel formulation of Auction Design as a two-player game with stationary utility functions.

A PAC-Bayesian Approach to Generalization Bounds for Graph Neural Networks

- Renjie Liao, Raquel Urtasun, Richard Zemel
- abstract@open-review(Poster): In this paper, we derive generalization bounds for two primary classes of graph neural networks (GNNs), namely graph convolutional networks (GCNs) and message passing GNNs (MPGNNs), via a PAC-Bayesian approach. Our result reveals that the maximum node degree and the spectral norm of the weights govern the generalization bounds of both models. We also show that our bound for GCNs is a natural generalization of the results developed in \citep{neyshabur2017pac} for fully-connected and convolutional neural networks. For MPGNNs, our PAC-Bayes bound improves over the Rademacher complexity based bound \citep{garg2020generalization}, showing a tighter dependency on the maximum node degree and the maximum hidden dimension. The key ingredients of our proofs are a perturbation analysis of GNNs and the generalization of PAC-Bayes analysis to non-homogeneous GNNs. We perform an empirical study on several synthetic and real-world graph datasets and verify that our PAC-Bayes bound is tighter than others.

Contextual Transformation Networks for Online Continual Learning

- Quang Pham, Chenghao Liu, Doyen Sahoo, Steven HOI
- abstract@open-review(Poster): Continual learning methods with fixed architectures rely on a single network to learn models that can perform well on all tasks. As a result, they often only accommodate common features of those tasks but neglect each task's specific features. On the other hand, dynamic architecture methods can have a separate network for each task, but they are too expensive to train and not scalable in practice, especially in online settings. To address this problem, we propose a novel online continual learning method named ``Contextual Transformation Networks'' (CTN) to efficiently model the \emph{task-specific features} while enjoying neglectable complexity overhead compared to other fixed architecture methods. Moreover, inspired by the Complementary Learning Systems (CLS) theory, we propose a novel dual memory design and an objective to train CTN that can address both catastrophic forgetting and knowledge transfer simultaneously. Our extensive experiments show that CTN is competitive with a large

scale dynamic architecture network and consistently outperforms other fixed architecture methods under the same standard backbone. Our implementation can be found at \url{https://github.com/phquang/Contextual-Transformation-Network}.

Disentangled Recurrent Wasserstein Autoencoder

- Jun Han, Martin Renqiang Min, Ligong Han, Li Erran Li, Xuan Zhang
- abstract@open-review(Spotlight): Learning disentangled representations leads to interpretable models and facilitates data generation with style transfer, which has been extensively studied on static data such as images in an unsupervised learning framework. However, only a few works have explored unsupervised disentangled sequential representation learning due to challenges of generating sequential data. In this paper, we propose recurrent Wasserstein Autoencoder (R-WAE), a new framework for generative modeling of sequential data. R-WAE disentangles the representation of an input sequence into static and dynamic factors (i.e., time-invariant and time-varying parts). Our theoretical analysis shows that, R-WAE minimizes an upper bound of a penalized form of the Wasserstein distance between model distribution and sequential data distribution, and simultaneously maximizes the mutual information between input data and different disentangled latent factors, respectively. This is superior to (recurrent) VAE which does not explicitly enforce mutual information maximization between input data and disentangled latent representations. When the number of actions in sequential data is available as weak supervision information, R-WAE is extended to learn a categorical latent representation of actions to improve its disentanglement. Experiments on a variety of datasets show that our models outperform other baselines with the same settings in terms of disentanglement and unconditional video generation both quantitatively and qualitatively.

Adaptive and Generative Zero-Shot Learning

- Yu-Ying Chou, Hsuan-Tien Lin, Tyng-Luh Liu
- abstract@open-review(Poster): We address the problem of generalized zero-shot learning (GZSL) where the task is to predict the class label of a target image whether its label belongs to the seen or unseen category. Similar to ZSL, the learning setting assumes that all class-level semantic features are given, while only the images of seen classes are available for training. By exploring the correlation between image features and the corresponding semantic features, the main idea of the proposed approach is to enrich the semantic-to-visual (S2V) embeddings via a seamless fusion of adaptive and generative learning. To this end, we extend the semantic features of each class by supplementing image-adaptive attention so that the learned S2V embedding can account for not only inter-class but also intra-class variations. In addition, to break the limit of training with images only from seen classes, we design a generative scheme to simultaneously generate virtual class labels and their visual features by sampling and interpolating over seen counterparts. In inference, a testing image will give rise to two different S2V embeddings, seen and virtual. The former is used to decide whether the underlying label is of the unseen category or otherwise a specific seen class; the latter is to predict an unseen class label. To demonstrate the effectiveness of our method, we report state-of-the-art results on four standard GZSL datasets, including an ablation study of the proposed modules.

Iterated learning for emergent systematicity in VQA

- Ankit Vani, Max Schwarzer, Yuchen Lu, Eeshan Dhekane, Aaron Courville
- abstract@open-review(Oral): Although neural module networks have an architectural bias towards compositionality, they require gold standard layouts to generalize systematically in practice. When instead learning layouts and modules jointly, compositionality does not arise automatically and an explicit pressure is necessary for the emergence of layouts exhibiting the right structure. We propose to address this problem using iterated learning, a cognitive science theory of the emergence of compositional languages in nature that has primarily been applied to simple referential games in machine learning. Considering the layouts of module networks as samples from an emergent language, we use iterated learning to encourage the development of structure within this language. We show that the resulting layouts support systematic generalization in neural agents solving the more complex task of visual question-answering. Our regularized iterated learning method can outperform baselines without iterated learning on SHAPES-SyGeT (SHAPES Systematic Generalization Test), a new split of the SHAPES dataset we introduce to evaluate systematic generalization, and on CLOSURE, an extension of CLEVR also designed to test systematic generalization. We demonstrate superior performance in recovering ground-truth compositional program structure with limited supervision on both SHAPES-SyGeT and CLEVR.

Online Adversarial Purification based on Self-supervised Learning

- Changhao Shi, Chester Holtz, Gal Mishne
- abstract@open-review(Poster): Deep neural networks are known to be vulnerable to adversarial examples, where a perturbation in the input space leads to an amplified shift in the latent network representation. In this paper, we combine canonical supervised learning with self-supervised representation learning, and present Self-supervised Online Adversarial Purification (SOAP), a novel defense strategy that uses a self-supervised loss to purify adversarial examples at test-time. Our approach leverages the label-independent nature of self-supervised signals and counters the adversarial perturbation with respect to the self-supervised tasks. SOAP yields competitive robust accuracy against state-of-the-art adversarial training and purification methods, with considerably less training complexity. In addition, our approach is robust even when adversaries are given the knowledge of the purification defense strategy. To the best of our knowledge, our paper is the first that generalizes the idea of using self-supervised signals to perform online test-time purification.

FairFil: Contrastive Neural Debiasing Method for Pretrained Text Encoders

- Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, Lawrence Carin
- abstract@open-review(Poster): Pretrained text encoders, such as BERT, have been applied increasingly in various natural language processing (NLP) tasks, and have recently demonstrated significant performance gains. However, recent studies have demonstrated the existence of social bias in these pretrained NLP models. Although prior works have made progress on word-level debiasing, improved sentence-level fairness of pretrained encoders still lacks exploration. In this paper, we proposed the first neural debiasing method for a pretrained sentence encoder, which transforms the pretrained encoder outputs into debiased representations via a fair filter (FairFil) network. To learn the FairFil, we introduce a contrastive learning framework that not only minimizes the correlation between filtered embeddings and bias words but also preserves rich semantic information of the original sentences. On real-world datasets, our FairFil effectively reduces the bias degree of pretrained text encoders, while continuously showing desirable performance on downstream tasks. Moreover, our post hoc method does not require any retraining of the text encoders, further enlarging FairFil's application space.

Reset-Free Lifelong Learning with Skill-Space Planning

- Kevin Lu, Aditya Grover, Pieter Abbeel, Igor Mordatch
- abstract@open-review(Poster): The objective of \textit{lifelong} reinforcement learning (RL) is to optimize agents which can continuously adapt and interact in changing environments. However, current RL approaches fail drastically when environments are non-stationary and interactions are non-episodic. We propose \textit{Lifelong Skill Planning} (LiSP), an algorithmic framework for lifelong RL based on planning in an abstract space of higher-order skills. We learn the skills in an unsupervised manner using intrinsic rewards and plan over the learned skills using a learned dynamics model. Moreover, our framework permits skill discovery even from offline data, thereby reducing the need for excessive real-world interactions. We demonstrate empirically that LiSP successfully enables long-horizon planning and learns agents that can avoid catastrophic failures even in challenging non-stationary and non-episodic environments derived from gridworld and MuJoCo benchmarks.

Efficient Empowerment Estimation for Unsupervised Stabilization

- Ruihan Zhao, Kevin Lu, Pieter Abbeel, Stas Tiomkin
- abstract@open-review(Poster): Intrinsically motivated artificial agents learn advantageous behavior without externally-provided rewards. Previously, it was shown that maximizing mutual information between agent actuators and future states, known as the empowerment principle, enables unsupervised stabilization of dynamical systems at upright positions, which is a prototypical intrinsically motivated behavior for upright standing and walking. This follows from the coincidence between the objective of stabilization and the objective of empowerment. Unfortunately, sample-based estimation of this kind of mutual information is challenging. Recently, various variational lower bounds (VLBs) on empowerment have been proposed as solutions; however, they are often biased, unstable in training, and have high sample complexity. In this work, we propose an alternative solution based on a trainable representation of a dynamical system as a Gaussian channel, which allows us to efficiently calculate an unbiased estimator of empowerment by convex optimization. We demonstrate our solution for sample-based unsupervised stabilization on different dynamical control systems and show the advantages of our method by comparing it to the existing VLB approaches. Specifically, we show that our method has a lower sample complexity, is more stable in training, possesses the essential properties of the empowerment function, and allows estimation of empowerment from images. Consequently, our method opens a path to wider and easier adoption of empowerment for various applications.

MixKD: Towards Efficient Distillation of Large-scale Language Models

- Kevin J Liang, Weituo Hao, Dinghan Shen, Yufan Zhou, Weizhu Chen, Changyou Chen, Lawrence Carin
- abstract@open-review(Poster): Large-scale language models have recently demonstrated impressive empirical performance. Nevertheless, the improved results are attained at the price of bigger models, more power consumption, and slower inference, which hinder their applicability to low-resource (both memory and computation) platforms. Knowledge distillation (KD) has been demonstrated as an effective framework for compressing such big models. However, large-scale neural network systems are prone to memorize training instances, and thus tend to make inconsistent predictions when the data distribution is altered slightly. Moreover, the student model has few opportunities to request useful information from the teacher model when there is limited task-specific data available. To address these issues, we propose MixKD, a data-agnostic distillation framework that leverages mixup, a simple yet efficient data augmentation approach, to endow the resulting model with stronger generalization ability. Concretely, in addition to the original training examples, the student model is encouraged to mimic the teacher's behavior on the linear interpolation of example pairs as well. We prove from a theoretical perspective that under reasonable conditions MixKD gives rise to a smaller gap between the generalization error and the empirical error. To verify its effectiveness, we conduct experiments on the GLUE benchmark, where MixKD consistently leads to significant gains over the standard KD training, and outperforms several competitive baselines. Experiments under a limited-data setting and ablation studies further demonstrate the advantages of the proposed approach.

CaPC Learning: Confidential and Private Collaborative Learning

- Christopher A. Choquette-Choo, Natalie Dullerud, Adam Dziedzic, Yunxiang Zhang, Somesh Jha, Nicolas Papernot, Xiao Wang
- abstract@open-review(Poster): Machine learning benefits from large training datasets, which may not always be possible to collect by any single entity, especially when using privacy-sensitive data. In many contexts, such as healthcare and finance, separate parties may wish to collaborate and learn from each other's data but are prevented from doing so due to privacy regulations. Some regulations prevent explicit sharing of data between parties by joining datasets in a central location (confidentiality). Others also limit implicit sharing of data, e.g., through model predictions (privacy). There is currently no method that enables machine learning in such a setting, where both confidentiality and privacy need to be preserved, to prevent both explicit and implicit sharing of data. Federated learning only provides confidentiality, not privacy, since gradients shared still contain private information. Differentially private learning assumes unreasonably large datasets. Furthermore, both of these learning paradigms produce a central model whose architecture was previously agreed upon by all parties rather than enabling collaborative learning where each party learns and improves their own local model. We introduce Confidential and Private Collaborative (CaPC) learning, the first method provably achieving both confidentiality and privacy in a collaborative setting. We leverage secure multi-party computation (MPC), homomorphic encryption (HE), and other techniques in combination with privately aggregated teacher models. We demonstrate how CaPC allows participants to collaborate without having to explicitly join their training sets or train a central model. Each party is able to improve the accuracy and fairness of their model, even in settings where each party has a model that performs well on their own dataset or when datasets are not IID and model architectures are heterogeneous across parties.

Multiplicative Filter Networks

- Rizal Fathony, Anit Kumar Sahu, Devin Willmott, J Zico Kolter
- abstract@open-review(Poster): Although deep networks are typically used to approximate functions over high dimensional inputs, recent work has increased interest in neural networks as function approximators for low-dimensional-but-complex functions, such as representing images as a function of pixel coordinates, solving differential equations, or representing signed distance fields or neural radiance fields. Key to these recent successes has been the use of new elements such as sinusoidal nonlinearities, or Fourier features in positional encodings, which vastly outperform simple ReLU networks. In this paper, we propose and empirically demonstrate that an arguably simpler class of function approximators can work just as well for such problems: multiplicative filter networks. In these networks, we avoid traditional compositional depth altogether, and simply multiply together (linear functions of) sinusoidal or Gabor wavelet functions applied to the input. This representation has the notable advantage that the entire function can simply be viewed as a linear function approximator over an exponential number of Fourier or Gabor basis functions, respectively. Despite this simplicity, when compared to recent approaches that use Fourier features with ReLU networks or sinusoidal activation networks, we show that these multiplicative filter networks largely outperform or match the performance of these recent approaches on the domains highlighted in these past works.

Planning from Pixels using Inverse Dynamics Models

- Keiran Paster, Sheila A. McIlraith, Jimmy Ba
- abstract@open-review(Poster): Learning dynamics models in high-dimensional observation spaces can be challenging for model-based RL agents. We propose a novel way to learn models in a latent space by learning to predict sequences of future actions conditioned on task completion. These models track task-relevant environment dynamics over a distribution of tasks, while simultaneously serving as an effective heuristic for planning with sparse rewards. We evaluate our method on challenging visual goal completion tasks and show a substantial increase in performance compared to prior model-free approaches.

Semi-supervised Keypoint Localization

- Olga Moskvyak, Frederic Maire, Feras Dayoub, Mahsa Baktashmotlagh
- abstract@open-review(Poster): Knowledge about the locations of keypoints of an object in an image can assist in fine-grained classification and identification tasks, particularly for the case of objects that exhibit large variations in poses that greatly influence their visual appearance, such as wild animals. However, supervised training of a keypoint detection network requires annotating a large image dataset for each animal species, which is a labor-intensive task. To reduce the need for labeled data, we propose to learn simultaneously keypoint heatmaps and pose invariant keypoint representations in a semi-supervised manner using a small set of labeled images along with a larger set of unlabeled images. Keypoint representations are learnt with a semantic keypoint consistency constraint that forces the keypoint detection network to learn similar features for the same keypoint across the dataset. Pose invariance is achieved by making keypoint representations for the image and its augmented copies closer together in feature space. Our semi-supervised approach significantly outperforms previous methods on several benchmarks for human and animal body landmark localization.

Influence Estimation for Generative Adversarial Networks

- Naoyuki Terashita, Hiroki Ohashi, Yuichi Nonaka, Takashi Kanemaru
- abstract@open-review(Spotlight): Identifying harmful instances, whose absence in a training dataset improves model performance, is important for building better machine learning models. Although previous studies have succeeded in estimating harmful instances under supervised settings, they cannot be trivially extended to generative adversarial networks (GANs). This is because previous approaches require that (i) the absence of a training instance directly affects the loss value and that (ii) the change in the loss directly measures the harmfulness of the instance for the performance of a model. In GAN training, however, neither of the requirements is satisfied. This is because, (i) the generator's loss is not directly affected by the training instances as they are not part of the generator's training steps, and (ii) the values of GAN's losses normally do not capture the generative performance of a model. To this end, (i) we propose an influence estimation method that uses the Jacobian of the gradient of the generator's loss with respect to the discriminator's parameters (and vice versa) to trace how the absence of an instance in the discriminator's training affects the generator's parameters, and (ii) we propose a novel evaluation scheme, in which we assess harmfulness of each training instance on the basis of how GAN evaluation metric (e.g., inception score) is expected to change due to the removal of the instance. We experimentally verified that our influence estimation method correctly inferred the changes in GAN evaluation metrics. We also demonstrated that the removal of the identified harmful instances effectively improved the model's generative performance with respect to various GAN evaluation metrics.

Emergent Road Rules In Multi-Agent Driving Environments

- Avik Pal, Jonah Philion, Yuan-Hong Liao, Sanja Fidler
- abstract@open-review(Poster): For autonomous vehicles to safely share the road with human drivers, autonomous vehicles must abide by specific "road rules" that human drivers have agreed to follow. "Road rules" include rules that drivers are required to follow by law – such as the requirement that vehicles stop at red lights – as well as more subtle social rules – such as the implicit designation of fast lanes on the highway. In this paper, we provide empirical evidence that suggests that – instead of hard-coding road rules into self-driving algorithms – a scalable alternative may be to design multi-agent environments in which road rules emerge as optimal solutions to the problem of maximizing traffic flow. We analyze what ingredients in driving environments cause the emergence of these road rules and find that two crucial factors are noisy perception and agents' spatial density. We provide qualitative and quantitative evidence of the emergence of seven social driving behaviors, ranging from obeying traffic signals to following lanes, all of which emerge from training agents to drive quickly to destinations without colliding. Our results add empirical support for the social road rules that countries worldwide have agreed on for safe, efficient driving.

SSD: A Unified Framework for Self-Supervised Outlier Detection

- Vikash Sehwag, Mung Chiang, Prateek Mittal
- abstract@open-review(Poster): We ask the following question: what training information is required to design an effective outlier/out-of-distribution (OOD) detector, i.e., detecting samples that lie far away from training distribution? Since unlabeled data is easily accessible for many applications, the most compelling approach is to develop detectors based on only unlabeled in-distribution data. However, we observe that most existing detectors based on unlabeled data perform poorly, often equivalent to a random prediction. In contrast, existing state-of-the-art OOD detectors achieve impressive performance but require access to fine-grained data labels for supervised training. We propose SSD, an outlier detector based on only unlabeled in-distribution data. We use self-supervised representation learning followed by a Mahalanobis distance based detection in the feature space. We demonstrate that SSD outperforms most existing detectors based on unlabeled data by a large margin. Additionally, SSD even achieves performance on par, and sometimes even better, with supervised training based detectors. Finally, we expand our detection framework with two key extensions. First, we formulate few-shot OOD detection, in which the detector has access to only one to five samples from each class of the targeted OOD dataset. Second, we extend our framework to incorporate training data labels, if available. We find that our novel detection framework based on SSD displays enhanced performance with these extensions, and achieves state-of-the-art performance. Our code is publicly available at https://github.com/inspire-group/SSD.

ECONOMIC HYPERPARAMETER OPTIMIZATION WITH BLENDED SEARCH STRATEGY

- Chi Wang, Qingyun Wu, Silu Huang, Amin Saied
- abstract@open-review(Poster): We study the problem of using low cost to search for hyperparameter configurations in a large search space with heterogeneous evaluation cost and model quality. We propose a blended search strategy to combine the strengths of global and local search, and prioritize them on the fly with the goal of minimizing the total cost spent in finding good configurations. Our approach demonstrates robust performance for tuning both tree-based models and deep neural networks on a large AutoML benchmark, as well as superior performance in model quality, time, and resource consumption for a production transformer-based NLP model fine-tuning task.

When Do Curricula Work?

- Xiaoxia Wu, Ethan Dyer, Behnam Neyshabur
- abstract@open-review(Oral): Inspired by human learning, researchers have proposed ordering examples during training based on their difficulty. Both curriculum learning, exposing a network to easier examples early in training, and anti-curriculum learning, showing the most difficult examples first, have been suggested as improvements to the standard i.i.d. training. In this work, we set out to investigate the relative benefits of ordered learning. We first investigate the implicit curricula resulting from architectural and optimization bias and find that samples are learned in a highly consistent order. Next, to quantify the benefit of explicit curricula, we conduct extensive experiments over thousands of orderings spanning three kinds of learning: curriculum, anti-curriculum, and random-curriculum -- in which the size of the training dataset is dynamically increased over time, but the examples are randomly ordered. We find that for standard benchmark datasets, curricula have only marginal benefits, and that randomly ordered samples perform as well or better than curricula and anti-curricula, suggesting that any benefit is entirely due to the dynamic training set size. Inspired by common use cases of curriculum learning in practice, we investigate the role of limited training time budget and noisy data in the success of curriculum learning. Our experiments demonstrate that curriculum, but not anti-curriculum or random ordering can indeed improve the performance either with limited training time budget or in the existence of noisy data.

Orthogonalizing Convolutional Layers with the Cayley Transform

- Asher Trockman, J Zico Kolter
- abstract@open-review(Spotlight): Recent work has highlighted several advantages of enforcing orthogonality in the weight layers of deep networks, such as maintaining the stability of activations, preserving gradient norms, and enhancing adversarial robustness by enforcing low Lipschitz constants. Although numerous methods exist for enforcing the orthogonality of fully-connected layers, those for convolutional layers are more heuristic in nature, often focusing on penalty methods or limited classes of convolutions. In this work, we propose and evaluate an alternative approach to directly parameterize convolutional layers that are constrained to be orthogonal. Specifically, we propose to apply the Cayley transform to a skew-symmetric convolution in the Fourier domain, so that the inverse convolution needed by the Cayley transform can be computed efficiently. We compare our method to previous Lipschitz-constrained and orthogonal convolutional layers and show that it indeed preserves orthogonality to a high degree even for large convolutions. Applied to the problem of certified adversarial robustness, we show that networks incorporating the layer outperform existing deterministic methods for certified defense against ℓ_2-norm-bounded adversaries, while scaling to larger architectures than previously investigated. Code is available at https://github.com/locuslab/orthogonal-convolutions.

Inductive Representation Learning in Temporal Networks via Causal Anonymous Walks

- Yanbang Wang, Yen-Yu Chang, Yunyu Liu, Jure Leskovec, Pan Li
- abstract@open-review(Poster): Temporal networks serve as abstractions of many real-world dynamic systems. These networks typically evolve according to certain laws, such as the law of triadic closure, which is universal in social networks. Inductive representation learning of temporal networks should be able to capture such laws and further be applied to systems that follow the same laws but have not been unseen during the training stage. Previous works in this area depend on either network node identities or rich edge attributes and typically fail to extract these laws. Here, we propose {\em Causal Anonymous Walks (CAWs)} to inductively represent a temporal network. CAWs are extracted by temporal random walks and work as automatic retrieval of temporal network motifs to represent network dynamics while avoiding the time-consuming selection and counting of those motifs. CAWs adopt a novel anonymization strategy that replaces node identities with the hitting counts of the nodes based on a set of sampled walks to keep the method inductive, and simultaneously establish the correlation between motifs. We further propose a neural-network model CAW-N to encode CAWs, and pair it with a CAW sampling strategy with constant memory and time cost to support online training and inference. CAW-N is evaluated to predict links over 6 real temporal networks and uniformly outperforms previous SOTA methods by averaged 15\% AUC gain in the inductive setting. CAW-N also outperforms previous methods in 5 out of the 6 networks in the transductive setting.

Robust Overfitting may be mitigated by properly learned smoothening

- Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, Zhangyang Wang
- abstract@open-review(Poster): A recent study (Rice et al., 2020) revealed overfitting to be a dominant phenomenon in adversarially robust training of deep networks, and that appropriate early-stopping of adversarial training (AT) could match the performance gains of most recent algorithmic improvements. This intriguing problem of robust overfitting motivates us to seek more remedies. As a pilot study, this paper investigates two empirical means to inject more learned smoothening during AT: one leveraging knowledge distillation and self-training to smooth the logits, the other performing stochastic weight averaging (Izmailov et al., 2018) to smooth the weights. Despite the embarrassing simplicity, the two approaches are surprisingly effective and hassle-free in mitigating robust overfitting. Experiments demonstrate that by plugging in them to AT, we can simultaneously boost the standard accuracy by $3.72\%\sim6.68\%$ and robust accuracy by $0.22\%\sim2 .03\%$, across multiple datasets (STL-10, SVHN, CIFAR-10, CIFAR-100, and Tiny ImageNet), perturbation types (ℓ_{∞} and ℓ_2), and robustified methods (PGD, TRADES, and FSGM), establishing the new state-of-the-art bar in AT. We present systematic visualizations and analyses to dive into their possible working mechanisms. We also carefully exclude the possibility of gradient masking by evaluating our models' robustness against transfer attacks. Codes are available at https://github.com/VITA-Group/Alleviate-Robust-Overfitting.

Predicting Infectiousness for Proactive Contact Tracing

- Yoshua Bengio, Prateek Gupta, Tegan Maharaj, Nasim Rahaman, Martin Weiss, Tristan Deleu, Eilif Benjamin Muller, Meng Qu, victor schmidt, Pierre-luc St-charles, hannah alsdurf, Olexa Bilaniuk, david buckeridge, gaetan caron, pierre luc carrier, Joumana Ghosn, satya ortiz gagne, Christopher Pal, Irina Rish, Bernhard Schölkopf, abhinav sharma, Jian Tang, andrew williams
- abstract@open-review(Spotlight): The COVID-19 pandemic has spread rapidly worldwide, overwhelming manual contact tracing in many countries and resulting in widespread lockdowns for emergency containment. Large-scale digital contact tracing (DCT) has emerged as a potential solution to resume economic and social activity while minimizing spread of the virus. Various DCT methods have been proposed, each making trade-offs be-tween privacy, mobility restrictions, and public health. The most common approach, binary contact tracing (BCT), models infection as a binary event, informed only by an individual's test results, with corresponding binary recommendations that either all or none of the individual's contacts quarantine. BCT ignores the inherent uncertainty in contacts and the infection process, which could be used to tailor messaging to high-risk individuals, and prompt proactive testing or earlier warnings. It also does not make use of observations such as symptoms or pre-existing medical conditions, which could be used to make more accurate infectiousness predictions. In this paper, we use a recently-proposed COVID-19 epidemiological simulator to develop and test methods that can be deployed to a smartphone to locally and proactively predict an individual's infectiousness (risk of infecting others) based on their contact history and other information, while respecting strong privacy constraints. Predictions are used to provide personalized recommendations to the individual via an app, as well as to send anonymized messages to the individual's contacts, who use this information to better predict their own infectiousness, an approach we call proactive contact tracing (PCT). Similarly to other works, we find that compared to no tracing, all DCT methods tested are able to reduce spread of the disease and thus save lives, even at low adoption rates, strongly supporting a role for DCT methods in managing the pandemic. Further, we find a deep-learning based PCT method which improves over BCT for equivalent average mobility, suggesting PCT could help in safe re-opening and second-wave prevention.

Local Search Algorithms for Rank-Constrained Convex Optimization

- Kyriakos Axiotis, Maxim Sviridenko
- abstract@open-review(Poster): We propose greedy and local search algorithms for rank-constrained convex optimization, namely solving $\underset{\mathrm{rank}(A)\leq r^{\}{\min}\, R(A)$ *given a convex function $R:\mathbb{R}^{m\times n}\rightarrow \mathbb{R}$ and a parameter $r^$.* These algorithms consist of repeating two steps: (a) adding a new rank-1 matrix to A and (b) enforcing the rank constraint on A. We refine and improve the theoretical analysis of Shalev-Shwartz et al. (2011), and show that if the rank-restricted condition number of R is κ, a solution A with rank $O(r^\cdot \min{\kappa \log \frac{R(\mathbf{0})-R(A^)}{\epsilon}, \kappa^2})$ and $R(A) \leq R(A^) + \epsilon$ *can be recovered, where $A^$* is the optimal solution. This significantly generalizes associated results on sparse convex optimization, as well as rank-constrained convex optimization for smooth functions. We then introduce new practical variants of these algorithms that have superior runtime and recover better solutions in practice. We demonstrate the versatility of these methods on a wide range of applications involving matrix completion and robust principal component analysis.

Learning Task Decomposition with Ordered Memory Policy Network

- Yuchen Lu, Yikang Shen, Siyuan Zhou, Aaron Courville, Joshua B. Tenenbaum, Chuang Gan
- abstract@open-review(Poster): Many complex real-world tasks are composed of several levels of subtasks. Humans leverage these hierarchical structures to accelerate the learning process and achieve better generalization. In this work, we study the inductive bias and propose Ordered Memory Policy Network (OMPN) to discover subtask hierarchy by learning from demonstration. The discovered subtask hierarchy could be used to perform task decomposition, recovering the subtask boundaries in an unstructured demonstration. Experiments on Craft and Dial demonstrate that our model can achieve higher task decomposition performance under both unsupervised and weakly supervised settings, comparing with strong baselines. OMPN can also be directly applied to partially observable environments and still achieve higher task decomposition performance. Our visualization further confirms that the subtask hierarchy can emerge in our model 1.

Property Controllable Variational Autoencoder via Invertible Mutual Dependence

- Xiaojie Guo, Yuanqi Du, Liang Zhao
- abstract@open-review(Poster): Deep generative models have made important progress towards modeling complex, high dimensional data via learning latent representations. Their usefulness is nevertheless often limited by a lack of control over the generative process or a poor understanding of the latent representation. To overcome these issues, attention is now focused on discovering latent variables correlated to the data properties and ways to manipulate these properties. This paper presents the new Property controllable VAE (PCVAE), where a new Bayesian model is proposed to inductively bias the latent

representation using explicit data properties via novel group-wise and property-wise disentanglement. Each data property corresponds seamlessly to a latent variable, by innovatively enforcing invertible mutual dependence between them. This allows us to move along the learned latent dimensions to control specific properties of the generated data with great precision. Quantitative and qualitative evaluations confirm that the PCVAE outperforms the existing models by up to 28% in capturing and 65% in manipulating the desired properties.

Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients

- Brenden K Petersen, Mikel Landajuela Larma, Terrell N. Mundhenk, Claudio Prata Santiago, Soo Kyung Kim, Joanne Taery Kim
- abstract@open-review(Oral): Discovering the underlying mathematical expressions describing a dataset is a core challenge for artificial intelligence. This is the problem of $\textit{symbolic regression}$. Despite recent advances in training neural networks to solve complex tasks, deep learning approaches to symbolic regression are underexplored. We propose a framework that leverages deep learning for symbolic regression via a simple idea: use a large model to search the space of small models. Specifically, we use a recurrent neural network to emit a distribution over tractable mathematical expressions and employ a novel risk-seeking policy gradient to train the network to generate better-fitting expressions. Our algorithm outperforms several baseline methods (including Eureqa, the gold standard for symbolic regression) in its ability to exactly recover symbolic expressions on a series of benchmark problems, both with and without added noise. More broadly, our contributions include a framework that can be applied to optimize hierarchical, variable-length objects under a black-box performance metric, with the ability to incorporate constraints in situ, and a risk-seeking policy gradient formulation that optimizes for best-case performance instead of expected performance.

Grounding Physical Concepts of Objects and Events Through Dynamic Visual Reasoning

- Zhenfang Chen, Jiayuan Mao, Jiajun Wu, Kwan-Yee Kenneth Wong, Joshua B. Tenenbaum, Chuang Gan
- abstract@open-review(Poster): We study the problem of dynamic visual reasoning on raw videos. This is a challenging problem; currently, state-of-the-art models often require dense supervision on physical object properties and events from simulation, which are impractical to obtain in real life. In this paper, we present the Dynamic Concept Learner (DCL), a unified framework that grounds physical objects and events from video and language. DCL first adopts a trajectory extractor to track each object over time and to represent it as a latent, object-centric feature vector. Building upon this object-centric representation, DCL learns to approximate the dynamic interaction among objects using graph networks. DCL further incorporates a semantic parser to parse question into semantic programs and, finally, a program executor to run the program to answer the question, levering the learned dynamics model. After training, DCL can detect and associate objects across the frames, ground visual properties and physical events, understand the causal relationship between events, make future and counterfactual predictions, and leverage these extracted presentations for answering queries. DCL achieves state-of-the-art performance on CLEVRER, a challenging causal video reasoning dataset, even without using ground-truth attributes and collision labels from simulations for training. We further test DCL on a newly proposed video-retrieval and event localization dataset derived from CLEVRER, showing its strong generalization capacity.

Learning a Latent Simplex in Input Sparsity Time

- Ainesh Bakshi, Chiranjib Bhattacharyya, Ravi Kannan, David Woodruff, Samson Zhou
- abstract@open-review(Spotlight): We consider the problem of learning a latent k-vertex simplex $K\in\mathbb{R}^d$, given $\mathbf{A}\in\mathbb{R}^{d\times n}$, which can be viewed as n data points that are formed by randomly perturbing some latent points in K, possibly beyond K. A large class of latent variable models, such as adversarial clustering, mixed membership stochastic block models, and topic models can be cast in this view of learning a latent simplex. Bhattacharyya and Kannan (SODA 2020) give an algorithm for learning such a k-vertex latent simplex in time roughly $O(k\cdot\text{nnz}(\mathbf{A}))$, where $\text{nnz}(\mathbf{A})$ is the number of non-zeros in $\mathbf{A}$. We show that the dependence on k in the running time is unnecessary given a natural assumption about the mass of the top k singular values of $\mathbf{A}$, which holds in many of these applications. Further, we show this assumption is necessary, as otherwise an algorithm for learning a latent simplex would imply a better low rank approximation algorithm than what is known.

We obtain a spectral low-rank approximation to $\mathbf{A}$ in input-sparsity time and show that the column space thus obtained has small $\sin\Theta$ (angular) distance to the right top-k singular space of $\mathbf{A}$. Our algorithm then selects k points in the low-rank subspace with the largest inner product (in absolute value) with k carefully chosen random vectors. By working in the low-rank subspace, we avoid reading the entire matrix in each iteration and thus circumvent the $\Theta(k\cdot\text{nnz}(\mathbf{A}))$ running time.

gradSim: Differentiable simulation for system identification and visuomotor control

- J. Krishna Murthy, Miles Macklin, Florian Golemo, Vikram Voleti, Linda Petrini, Martin Weiss, Breandan Considine, Jérôme Parent-Lévesque, Kevin Xie, Kenny Erleben, Liam Paull, Florian Shkurti, Derek Nowrouzezahrai, Sanja Fidler
- abstract@open-review(Poster): In this paper, we tackle the problem of estimating object physical properties such as mass, friction, and elasticity directly from video sequences. Such a system identification problem is fundamentally ill-posed due to the loss of information during image formation. Current best solutions to the problem require precise 3D labels which are labor intensive to gather, and infeasible to create for many systems such as deformable solids or cloth. In this work we present gradSim, a framework that overcomes the dependence on 3D supervision by combining differentiable multiphysics simulation and differentiable rendering to jointly model the evolution of scene dynamics and image formation. This unique combination enables backpropagation from pixels in a video sequence through to the underlying physical attributes that generated them. Furthermore, our unified computation graph across dynamics and rendering engines enables the learning of challenging visuomotor control tasks, without relying on state-based (3D) supervision, while obtaining performance competitive to/better than techniques that require precise 3D labels.

Generative Scene Graph Networks

- Fei Deng, Zhuo Zhi, Donghun Lee, Sungjin Ahn
- abstract@open-review(Poster): Human perception excels at building compositional hierarchies of parts and objects from unlabeled scenes that help systematic generalization. Yet most work on generative scene modeling either ignores the part-whole relationship or assumes access to predefined part labels. In this paper, we propose Generative Scene Graph Networks (GSGNs), the first deep generative model that learns to discover the primitive parts and infer the part-whole relationship jointly from multi-object scenes without supervision and in an end-to-end trainable way. We formulate GSGN as a variational autoencoder in which the latent representation is a tree-structured probabilistic scene graph. The leaf nodes in the latent tree correspond to primitive parts, and the edges represent the symbolic pose variables required for recursively composing the parts into whole objects and then the full scene. This allows novel objects and scenes to be generated both by sampling from the prior and by manual configuration of the pose variables, as we do with graphics engines. We evaluate GSGN on datasets of scenes containing multiple compositional objects, including a challenging Compositional CLEVR dataset that we have developed. We show that GSGN is able to infer the latent scene graph, generalize out of the training regime, and improve data efficiency in downstream tasks.

Decentralized Attribution of Generative Models

- Changhoon Kim, Yi Ren, Yezhou Yang
- abstract@open-review(Poster): Growing applications of generative models have led to new threats such as malicious personation and digital copyright infringement. One solution to these threats is model attribution, i.e., the identification of user-end models where the contents under question are generated.

Existing studies showed empirical feasibility of attribution through a centralized classifier trained on all existing user-end models. However, this approach is not scalable in a reality where the number of models ever grows. Neither does it provide an attributability guarantee. To this end, this paper studies decentralized attribution, which relies on binary classifiers associated with each user-end model. Each binary classifier is parameterized by a user-specific key and distinguishes its associated model distribution from the authentic data distribution. We develop sufficient conditions of the keys that guarantee an attributability lower bound. Our method is validated on MNIST, CelebA, and FFHQ datasets. We also examine the trade-off between generation quality and robustness of attribution against adversarial post-processes.

Improved Autoregressive Modeling with Distribution Smoothing

- Chenlin Meng, Jiaming Song, Yang Song, Shengjia Zhao, Stefano Ermon
- abstract@open-review(Oral): While autoregressive models excel at image compression, their sample quality is often lacking. Although not realistic, generated images often have high likelihood according to the model, resembling the case of adversarial examples. Inspired by a successful adversarial defense method, we incorporate randomized smoothing into autoregressive generative modeling. We first model a smoothed version of the data distribution, and then reverse the smoothing process to recover the original data distribution. This procedure drastically improves the sample quality of existing autoregressive models on several synthetic and real-world image datasets while obtaining competitive likelihoods on synthetic datasets.

CPT: Efficient Deep Neural Network Training via Cyclic Precision

- Yonggan Fu, Han Guo, Meng Li, Xin Yang, Yining Ding, Vikas Chandra, Yingyan Lin
- abstract@open-review(Spotlight): Low-precision deep neural network (DNN) training has gained tremendous attention as reducing precision is one of the most effective knobs for boosting DNNs' training time/energy efficiency. In this paper, we attempt to explore low-precision training from a new perspective as inspired by recent findings in understanding DNN training: we conjecture that DNNs' precision might have a similar effect as the learning rate during DNN training, and advocate dynamic precision along the training trajectory for further boosting the time/energy efficiency of DNN training. Specifically, we propose Cyclic Precision Training (CPT) to cyclically vary the precision between two boundary values which can be identified using a simple precision range test within the first few training epochs. Extensive simulations and ablation studies on five datasets and eleven models demonstrate that CPT's effectiveness is consistent across various models/tasks (including classification and language modeling). Furthermore, through experiments and visualization we show that CPT helps to (1) converge to a wider minima with a lower generalization error and (2) reduce training variance which we believe opens up a new design knob for simultaneously improving the optimization and efficiency of DNN training.

Individually Fair Rankings

- Amanda Bower, Hamid Eftekhari, Mikhail Yurochkin, Yuekai Sun
- abstract@open-review(Poster): We develop an algorithm to train individually fair learning-to-rank (LTR) models. The proposed approach ensures items from minority groups appear alongside similar items from majority groups. This notion of fair ranking is based on the definition of individual fairness from supervised learning and is more nuanced than prior fair LTR approaches that simply ensure the ranking model provides underrepresented items with a basic level of exposure. The crux of our method is an optimal transport-based regularizer that enforces individual fairness and an efficient algorithm for optimizing the regularizer. We show that our approach leads to certifiably individually fair LTR models and demonstrate the efficacy of our method on ranking tasks subject to demographic biases.

Watch-And-Help: A Challenge for Social Perception and Human-AI Collaboration

- Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B. Tenenbaum, Sanja Fidler, Antonio Torralba
- abstract@open-review(Spotlight): In this paper, we introduce Watch-And-Help (WAH), a challenge for testing social intelligence in agents. In WAH, an AI agent needs to help a human-like agent perform a complex household task efficiently. To succeed, the AI agent needs to i) understand the underlying goal of the task by watching a single demonstration of the human-like agent performing the same task (social perception), and ii) coordinate with the human-like agent to solve the task in an unseen environment as fast as possible (human-AI collaboration). For this challenge, we build VirtualHome-Social, a multi-agent household environment, and provide a benchmark including both planning and learning based baselines. We evaluate the performance of AI agents with the human-like agent as well as and with real humans using objective metrics and subjective user ratings. Experimental results demonstrate that our challenge and virtual environment enable a systematic evaluation on the important aspects of machine social intelligence at scale.

Adaptive Federated Optimization

- Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, Hugh Brendan McMahan
- abstract@open-review(Poster): Federated learning is a distributed machine learning paradigm in which a large number of clients coordinate with a central server to learn a model without sharing their own training data. Standard federated optimization methods such as Federated Averaging (FedAvg) are often difficult to tune and exhibit unfavorable convergence behavior. In non-federated settings, adaptive optimization methods have had notable success in combating such issues. In this work, we propose federated versions of adaptive optimizers, including Adagrad, Adam, and Yogi, and analyze their convergence in the presence of heterogeneous data for general non-convex settings. Our results highlight the interplay between client heterogeneity and communication efficiency. We also perform extensive experiments on these methods and show that the use of adaptive optimizers can significantly improve the performance of federated learning.

GANs Can Play Lottery Tickets Too

- Xuxi Chen, Zhenyu Zhang, Yongduo Sui, Tianlong Chen
- abstract@open-review(Poster): Deep generative adversarial networks (GANs) have gained growing popularity in numerous scenarios, while usually suffer from high parameter complexities for resource-constrained real-world applications. However, the compression of GANs has less been explored. A few works show that heuristically applying compression techniques normally leads to unsatisfactory results, due to the notorious training instability of GANs. In parallel, the lottery ticket hypothesis shows prevailing success on discriminative models, in locating sparse matching subnetworks capable of training in isolation to full model performance. In this work, we for the first time study the existence of such trainable matching subnetworks in deep GANs. For a range of GANs, we certainly find matching subnetworks at 67%-74% sparsity. We observe that with or without pruning discriminator has a minor effect on the existence and quality of matching subnetworks, while the initialization weights used in the discriminator plays a significant role. We then show the powerful transferability of these subnetworks to unseen tasks. Furthermore, extensive experimental results demonstrate that our found subnetworks substantially outperform previous state-of-the-art GAN compression approaches in both image generation (e.g. SNGAN) and image-to-image translation GANs (e.g. CycleGAN). Codes available at https://github.com/VITA-Group/GAN-LTH.

Score-Based Generative Modeling through Stochastic Differential Equations

- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, Ben Poole
- abstract@open-review(Oral): Creating noise from data is easy; creating data from noise is generative modeling. We present a stochastic differential equation (SDE) that smoothly transforms a complex data distribution to a known prior distribution by slowly injecting noise, and a corresponding reverse-

time SDE that transforms the prior distribution back into the data distribution by slowly removing the noise. Crucially, the reverse-time SDE depends only on the time-dependent gradient field (a.k.a., score) of the perturbed data distribution. By leveraging advances in score-based generative modeling, we can accurately estimate these scores with neural networks, and use numerical SDE solvers to generate samples. We show that this framework encapsulates previous approaches in score-based generative modeling and diffusion probabilistic modeling, allowing for new sampling procedures and new modeling capabilities. In particular, we introduce a predictor-corrector framework to correct errors in the evolution of the discretized reverse-time SDE. We also derive an equivalent neural ODE that samples from the same distribution as the SDE, but additionally enables exact likelihood computation, and improved sampling efficiency. In addition, we provide a new way to solve inverse problems with score-based models, as demonstrated with experiments on class-conditional generation, image inpainting, and colorization. Combined with multiple architectural improvements, we achieve record-breaking performance for unconditional image generation on CIFAR-10 with an Inception score of 9.89 and FID of 2.20, a competitive likelihood of 2.99 bits/dim, and demonstrate high fidelity generation of 1024×1024 images for the first time from a score-based generative model.

Improving Relational Regularized Autoencoders with Spherical Sliced Fused Gromov Wasserstein

- Khai Nguyen, Son Nguyen, Nhat Ho, Tung Pham, Hung Bui
- abstract@open-review(Poster): Relational regularized autoencoder (RAE) is a framework to learn the distribution of data by minimizing a reconstruction loss together with a relational regularization on the prior of latent space. A recent attempt to reduce the inner discrepancy between the prior and aggregated posterior distributions is to incorporate sliced fused Gromov-Wasserstein (SFG) between these distributions. That approach has a weakness since it treats every slicing direction similarly, meanwhile several directions are not useful for the discriminative task. To improve the discrepancy and consequently the relational regularization, we propose a new relational discrepancy, named spherical sliced fused Gromov Wasserstein (SSFG), that can find an important area of projections characterized by a von Mises-Fisher distribution. Then, we introduce two variants of SSFG to improve its performance. The first variant, named mixture spherical sliced fused Gromov Wasserstein (MSSFG), replaces the vMF distribution by a mixture of von Mises-Fisher distributions to capture multiple important areas of directions that are far from each other. The second variant, named power spherical sliced fused Gromov Wasserstein (PSSFG), replaces the vMF distribution by a power spherical distribution to improve the sampling time of the vMF distribution in high dimension settings. We then apply the new discrepancies to the RAE framework to achieve its new variants. Finally, we conduct extensive experiments to show that the new autoencoders have favorable performance in learning latent manifold structure, image generation, and reconstruction.

Learning Reasoning Paths over Semantic Graphs for Video-grounded Dialogues

- Hung Le, Nancy F. Chen, Steven Hoi
- abstract@open-review(Poster): Compared to traditional visual question answering, video-grounded dialogues require additional reasoning over dialogue context to answer questions in a multi-turn setting. Previous approaches to video-grounded dialogues mostly use dialogue context as a simple text input without modelling the inherent information flows at the turn level. In this paper, we propose a novel framework of Reasoning Paths in Dialogue Context (PDC). PDC model discovers information flows among dialogue turns through a semantic graph constructed based on lexical components in each question and answer. PDC model then learns to predict reasoning paths over this semantic graph. Our path prediction model predicts a path from the current turn through past dialogue turns that contain additional visual cues to answer the current question. Our reasoning model sequentially processes both visual and textual information through this reasoning path and the propagated features are used to generate the answer. Our experimental results demonstrate the effectiveness of our method and provide additional insights on how models use semantic dependencies in a dialogue context to retrieve visual cues.

Extreme Memorization via Scale of Initialization

- Harsh Mehta, Ashok Cutkosky, Behnam Neyshabur
- abstract@open-review(Poster): We construct an experimental setup in which changing the scale of initialization strongly impacts the implicit regularization induced by SGD, interpolating from good generalization performance to completely memorizing the training set while making little progress on the test set. Moreover, we find that the extent and manner in which generalization ability is affected depends on the activation and loss function used, with sin activation being the most extreme. In the case of the homogeneous ReLU activation, we show that this behavior can be attributed to the loss function. Our empirical investigation reveals that increasing the scale of initialization correlates with misalignment of representations and gradients across examples in the same class. This insight allows us to device an alignment measure over gradients and representations which can capture this phenomenon. We demonstrate that our alignment measure correlates with generalization of deep models trained on image classification tasks.

Teaching with Commentaries

- Aniruddh Raghu, Maithra Raghu, Simon Kornblith, David Duvenaud, Geoffrey Hinton
- abstract@open-review(Poster): Effective training of deep neural networks can be challenging, and there remain many open questions on how to best learn these models. Recently developed methods to improve neural network training examine teaching: providing learned information during the training process to improve downstream model performance. In this paper, we take steps towards extending the scope of teaching. We propose a flexible teaching framework using commentaries, learned meta-information helpful for training on a particular task. We present gradient-based methods to learn commentaries, leveraging recent work on implicit differentiation for scalability. We explore diverse applications of commentaries, from weighting training examples, to parameterising label-dependent data augmentation policies, to representing attention masks that highlight salient image regions. We find that commentaries can improve training speed and/or performance, and provide insights about the dataset and training process. We also observe that commentaries generalise: they can be reused when training new models to obtain performance benefits, suggesting a use-case where commentaries are stored with a dataset and leveraged in future for improved model training.

Gradient Vaccine: Investigating and Improving Multi-task Optimization in Massively Multilingual Models

- Zirui Wang, Yulia Tsvetkov, Orhan Firat, Yuan Cao
- abstract@open-review(Spotlight): Massively multilingual models subsuming tens or even hundreds of languages pose great challenges to multi-task optimization. While it is a common practice to apply a language-agnostic procedure optimizing a joint multilingual task objective, how to properly characterize and take advantage of its underlying problem structure for improving optimization efficiency remains under-explored. In this paper, we attempt to peek into the black-box of multilingual optimization through the lens of loss function geometry. We find that gradient similarity measured along the optimization trajectory is an important signal, which correlates well with not only language proximity but also the overall model performance. Such observation helps us to identify a critical limitation of existing gradient-based multi-task learning methods, and thus we derive a simple and scalable optimization procedure, named Gradient Vaccine, which encourages more geometrically aligned parameter updates for close tasks. Empirically, our method obtains significant model performance gains on multilingual machine translation and XTREME benchmark tasks for multilingual language models. Our work reveals the importance of properly measuring and utilizing language proximity in multilingual optimization, and has broader implications for multi-task learning beyond multilingual modeling.

PlasticineLab: A Soft-Body Manipulation Benchmark with Differentiable Physics

- Zhiao Huang, Yuanming Hu, Tao Du, Siyuan Zhou, Hao Su, Joshua B. Tenenbaum, Chuang Gan
- abstract@open-review(Spotlight): Simulated virtual environments serve as one of the main driving forces behind developing and evaluating skill learning algorithms. However, existing environments typically only simulate rigid body physics. Additionally, the simulation process usually does not provide gradients that might be useful for planning and control optimizations. We introduce a new differentiable physics benchmark called PasticineLab, which

includes a diverse collection of soft body manipulation tasks. In each task, the agent uses manipulators to deform the plasticine into a desired configuration. The underlying physics engine supports differentiable elastic and plastic deformation using the DiffTaichi system, posing many under-explored challenges to robotic agents. We evaluate several existing reinforcement learning (RL) methods and gradient-based methods on this benchmark. Experimental results suggest that 1) RL-based approaches struggle to solve most of the tasks efficiently; 2) gradient-based approaches, by optimizing open-loop control sequences with the built-in differentiable physics engine, can rapidly find a solution within tens of iterations, but still fall short on multi-stage tasks that require long-term planning. We expect that PlasticineLab will encourage the development of novel algorithms that combine differentiable physics and RL for more complex physics-based skill learning tasks. PlasticineLab will be made publicly available.

In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning

- Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, Mubarak Shah
- abstract@open-review(Poster): The recent research in semi-supervised learning (SSL) is mostly dominated by consistency regularization based methods which achieve strong performance. However, they heavily rely on domain-specific data augmentations, which are not easy to generate for all data modalities. Pseudo-labeling (PL) is a general SSL approach that does not have this constraint but performs relatively poorly in its original formulation. We argue that PL underperforms due to the erroneous high confidence predictions from poorly calibrated models; these predictions generate many incorrect pseudo-labels, leading to noisy training. We propose an uncertainty-aware pseudo-label selection (UPS) framework which improves pseudo labeling accuracy by drastically reducing the amount of noise encountered in the training process. Furthermore, UPS generalizes the pseudo-labeling process, allowing for the creation of negative pseudo-labels; these negative pseudo-labels can be used for multi-label classification as well as negative learning to improve the single-label classification. We achieve strong performance when compared to recent SSL methods on the CIFAR-10 and CIFAR-100 datasets. Also, we demonstrate the versatility of our method on the video dataset UCF-101 and the multi-label dataset Pascal VOC.

Into the Wild with AudioScope: Unsupervised Audio-Visual Separation of On-Screen Sounds

- Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Dan Ellis, John R. Hershey
- abstract@open-review(Poster): Recent progress in deep learning has enabled many advances in sound separation and visual scene understanding. However, extracting sound sources which are apparent in natural videos remains an open problem. In this work, we present AudioScope, a novel audio-visual sound separation framework that can be trained without supervision to isolate on-screen sound sources from real in-the-wild videos. Prior audio-visual separation work assumed artificial limitations on the domain of sound classes (e.g., to speech or music), constrained the number of sources, and required strong sound separation or visual segmentation labels. AudioScope overcomes these limitations, operating on an open domain of sounds, with variable numbers of sources, and without labels or prior visual segmentation. The training procedure for AudioScope uses mixture invariant training (MixIT) to separate synthetic mixtures of mixtures (MoMs) into individual sources, where noisy labels for mixtures are provided by an unsupervised audio-visual coincidence model. Using the noisy labels, along with attention between video and audio features, AudioScope learns to identify audio-visual similarity and to suppress off-screen sounds. We demonstrate the effectiveness of our approach using a dataset of video clips extracted from open-domain YFCC100m video data. This dataset contains a wide diversity of sound classes recorded in unconstrained conditions, making the application of previous methods unsuitable. For evaluation and semi-supervised experiments, we collected human labels for presence of on-screen and off-screen sounds on a small subset of clips.

Cut out the annotator, keep the cutout: better segmentation with weak supervision

- Sarah Hooper, Michael Wornow, Ying Hang Seah, Peter Kellman, Hui Xue, Frederic Sala, Curtis Langlotz, Christopher Re
- abstract@open-review(Poster): Constructing large, labeled training datasets for segmentation models is an expensive and labor-intensive process. This is a common challenge in machine learning, addressed by methods that require few or no labeled data points such as few-shot learning (FSL) and weakly-supervised learning (WS). Such techniques, however, have limitations when applied to image segmentation---FSL methods often produce noisy results and are strongly dependent on which few datapoints are labeled, while WS models struggle to fully exploit rich image information. We propose a framework that fuses FSL and WS for segmentation tasks, enabling users to train high-performing segmentation networks with very few hand-labeled training points. We use FSL models as weak sources in a WS framework, requiring a very small set of reference labeled images, and introduce a new WS model that focuses on key areas---areas with contention among noisy labels---of the image to fuse these weak sources. Empirically, we evaluate our proposed approach over seven well-motivated segmentation tasks. We show that our methods can achieve within 1.4 Dice points compared to fully supervised networks while only requiring five hand-labeled training points. Compared to existing FSL methods, our approach improves performance by a mean 3.6 Dice points over the next-best method.

CoDA: Contrast-enhanced and Diversity-promoting Data Augmentation for Natural Language Understanding

- Yanru Qu, Dinghan Shen, Yelong Shen, Sandra Sajeev, Weizhu Chen, Jiawei Han
- abstract@open-review(Poster): Data augmentation has been demonstrated as an effective strategy for improving model generalization and data efficiency. However, due to the discrete nature of natural language, designing label-preserving transformations for text data tends to be more challenging. In this paper, we propose a novel data augmentation frame-work dubbed CoDA, which synthesizes diverse and informative augmented examples by integrating multiple transformations organically. Moreover, a contrastive regularization is introduced to capture the global relationship among all the data samples. A momentum encoder along with a memory bank is further leveraged to better estimate the contrastive loss. To verify the effectiveness of the proposed framework, we apply CoDA to Transformer-based models on a wide range of natural language understanding tasks. On the GLUE benchmark, CoDA gives rise to an average improvement of 2.2%while applied to the Roberta-large model. More importantly, it consistently exhibits stronger results relative to several competitive data augmentation and adversarial training baselines (including the low-resource settings). Extensive experiments show that the proposed contrastive objective can be flexibly combined with various data augmentation approaches to further boost their performance, highlighting the wide applicability of the CoDA framework.

Global Convergence of Three-layer Neural Networks in the Mean Field Regime

- Huy Tuan Pham, Phan-Minh Nguyen
- abstract@open-review(Oral): In the mean field regime, neural networks are appropriately scaled so that as the width tends to infinity, the learning dynamics tends to a nonlinear and nontrivial dynamical limit, known as the mean field limit. This lends a way to study large-width neural networks via analyzing the mean field limit. Recent works have successfully applied such analysis to two-layer networks and provided global convergence guarantees. The extension to multilayer ones however has been a highly challenging puzzle, and little is known about the optimization efficiency in the mean field regime when there are more than two layers.

In this work, we prove a global convergence result for unregularized feedforward three-layer networks in the mean field regime. We first develop a rigorous framework to establish the mean field limit of three-layer networks under stochastic gradient descent training. To that end, we propose the idea of a neuronal embedding, which comprises of a fixed probability space that encapsulates neural networks of arbitrary sizes. The identified mean field limit is then used to prove a global convergence guarantee under suitable regularity and convergence mode assumptions, which – unlike previous works on two-layer networks – does not rely critically on convexity. Underlying the result is a universal approximation property, natural of neural networks, which importantly is shown to hold at any finite training time (not necessarily at convergence) via an algebraic topology argument.

Deep Learning meets Projective Clustering

- Alaa Maalouf, Harry Lang, Daniela Rus, Dan Feldman
- abstract@open-review(Poster): A common approach for compressing Natural Language Processing (NLP) networks is to encode the embedding layer as a matrix $A\in\mathbb{R}^{n\times d}$, compute its rank-$j$ approximation A_j via SVD (Singular Value Decomposition), and then factor A_j into a pair of matrices that correspond to smaller fully-connected layers to replace the original embedding layer. Geometrically, the rows of A represent points in $\mathbb{R}^d$, and the rows of A_j represent their projections onto the j-dimensional subspace that minimizes the sum of squared distances (``errors'') to the points. In practice, these rows of A may be spread around $k>1$ subspaces, so factoring A based on a single subspace may lead to large errors that turn into large drops in accuracy.

Inspired by \emph{projective clustering} from computational geometry, we suggest replacing this subspace by a set of k subspaces, each of dimension j, that minimizes the sum of squared distances over every point (row in A) to its \emph{closest} subspace. Based on this approach, we provide a novel architecture that replaces the original embedding layer by a set of k small layers that operate in parallel and are then recombined with a single fully-connected layer.

Extensive experimental results on the GLUE benchmark yield networks that are both more accurate and smaller compared to the standard matrix factorization (SVD). For example, we further compress DistilBERT by reducing the size of the embedding layer by 40% while incurring only a 0.5% average drop in accuracy over all nine GLUE tasks, compared to a 2.8% drop using the existing SVD approach. On RoBERTa we achieve 43% compression of the embedding layer with less than a 0.8% average drop in accuracy as compared to a 3% drop previously.

Learning to Deceive Knowledge Graph Augmented Models via Targeted Perturbation

- Mrigank Raman, Aaron Chan, Siddhant Agarwal, PeiFeng Wang, Hansen Wang, Sungchul Kim, Ryan Rossi, Handong Zhao, Nedim Lipka, Xiang Ren
- abstract@open-review(Poster): Knowledge graphs (KGs) have helped neural models improve performance on various knowledge-intensive tasks, like question answering and item recommendation. By using attention over the KG, such KG-augmented models can also "explain" which KG information was most relevant for making a given prediction. In this paper, we question whether these models are really behaving as we expect. We show that, through a reinforcement learning policy (or even simple heuristics), one can produce deceptively perturbed KGs, which maintain the downstream performance of the original KG while significantly deviating from the original KG's semantics and structure. Our findings raise doubts about KG-augmented models' ability to reason about KG information and give sensible explanations.

Knowledge Distillation as Semiparametric Inference

- Tri Dao, Govinda M Kamath, Vasilis Syrgkanis, Lester Mackey
- abstract@open-review(Poster): A popular approach to model compression is to train an inexpensive student model to mimic the class probabilities of a highly accurate but cumbersome teacher model. Surprisingly, this two-step knowledge distillation process often leads to higher accuracy than training the student directly on labeled data. To explain and enhance this phenomenon, we cast knowledge distillation as a semiparametric inference problem with the optimal student model as the target, the unknown Bayes class probabilities as nuisance, and the teacher probabilities as a plug-in nuisance estimate. By adapting modern semiparametric tools, we derive new guarantees for the prediction error of standard distillation and develop two enhancements—cross-fitting and loss correction—to mitigate the impact of teacher overfitting and underfitting on student performance. We validate our findings empirically on both tabular and image data and observe consistent improvements from our knowledge distillation enhancements.

Meta-Learning with Neural Tangent Kernels

- Yufan Zhou, Zhenyi Wang, Jiayi Xian, Changyou Chen, Jinhui Xu
- abstract@open-review(Poster): Model Agnostic Meta-Learning (MAML) has emerged as a standard framework for meta-learning, where a meta-model is learned with the ability of fast adapting to new tasks. However, as a double-looped optimization problem, MAML needs to differentiate through the whole inner-loop optimization path for every outer-loop training step, which may lead to both computational inefficiency and sub-optimal solutions. In this paper, we generalize MAML to allow meta-learning to be defined in function spaces, and propose the first meta-learning paradigm in the Reproducing Kernel Hilbert Space (RKHS) induced by the meta-model's Neural Tangent Kernel (NTK). Within this paradigm, we introduce two meta-learning algorithms in the RKHS, which no longer need a sub-optimal iterative inner-loop adaptation as in the MAML framework. We achieve this goal by 1) replacing the adaptation with a fast-adaptive regularizer in the RKHS; and 2) solving the adaptation analytically based on the NTK theory. Extensive experimental studies demonstrate advantages of our paradigm in both efficiency and quality of solutions compared to related meta-learning algorithms. Another interesting feature of our proposed methods is that they are demonstrated to be more robust to adversarial attacks and out-of-distribution adaptation than popular baselines, as demonstrated in our experiments.

Vulnerability-Aware Poisoning Mechanism for Online RL with Unknown Dynamics

- Yanchao Sun, Da Huo, Furong Huang
- abstract@open-review(Poster): Poisoning attacks on Reinforcement Learning (RL) systems could take advantage of RL algorithm’s vulnerabilities and cause failure of the learning. However, prior works on poisoning RL usually either unrealistically assume the attacker knows the underlying Markov Decision Process (MDP), or directly apply the poisoning methods in supervised learning to RL. In this work, we build a generic poisoning framework for online RL via a comprehensive investigation of heterogeneous poisoning models in RL. Without any prior knowledge of the MDP, we propose a strategic poisoning algorithm called Vulnerability-Aware Adversarial Critic Poison (VA2C-P), which works for on-policy deep RL agents, closing the gap that no poisoning method exists for policy-based RL agents. VA2C-P uses a novel metric, stability radius in RL, that measures the vulnerability of RL algorithms. Experiments on multiple deep RL agents and multiple environments show that our poisoning algorithm successfully prevents agents from learning a good policy or teaches the agents to converge to a target policy, with a limited attacking budget.

Understanding and Improving Lexical Choice in Non-Autoregressive Translation

- Liang Ding, Longyue Wang, Xuebo Liu, Derek F. Wong, Dacheng Tao, Zhaopeng Tu
- abstract@open-review(Poster): Knowledge distillation (KD) is essential for training non-autoregressive translation (NAT) models by reducing the complexity of the raw data with an autoregressive teacher model. In this study, we empirically show that as a side effect of this training, the lexical choice errors on low-frequency words are propagated to the NAT model from the teacher model. To alleviate this problem, we propose to expose the raw data to NAT models to restore the useful information of low-frequency words, which are missed in the distilled data. To this end, we introduce an extra Kullback-Leibler divergence term derived by comparing the lexical choice of NAT model and that embedded in the raw data. Experimental results across language pairs and model architectures demonstrate the effectiveness and universality of the proposed approach. Extensive analyses confirm our claim that our approach improves performance by reducing the lexical choice errors on low-frequency words. Encouragingly, our approach pushes the SOTA NAT performance on the WMT14 English-German and WMT16 Romanian-English datasets up to 27.8 and 33.8 BLEU points, respectively.

Rethinking Architecture Selection in Differentiable NAS

- Ruochen Wang, Minhao Cheng, Xiangning Chen, Xiaocheng Tang, Cho-Jui Hsieh
- abstract@open-review(Oral): Differentiable Neural Architecture Search is one of the most popular Neural Architecture Search (NAS) methods for its search efficiency and simplicity, accomplished by jointly optimizing the model weight and architecture parameters in a weight-sharing supernet via

gradient-based algorithms. At the end of the search phase, the operations with the largest architecture parameters will be selected to form the final architecture, with the implicit assumption that the values of architecture parameters reflect the operation strength. While much has been discussed about the supernet's optimization, the architecture selection process has received little attention. We provide empirical and theoretical analysis to show that the magnitude of architecture parameters does not necessarily indicate how much the operation contributes to the supernet's performance. We propose an alternative perturbation-based architecture selection that directly measures each operation's influence on the supernet. We re-evaluate several differentiable NAS methods with the proposed architecture selection and find that it is able to extract significantly improved architectures from the underlying supernets consistently. Furthermore, we find that several failure modes of DARTS can be greatly alleviated with the proposed selection method, indicating that much of the poor generalization observed in DARTS can be attributed to the failure of magnitude-based architecture selection rather than entirely the optimization of its supernet.

Distributional Sliced-Wasserstein and Applications to Generative Modeling

- Khai Nguyen, Nhat Ho, Tung Pham, Hung Bui
- abstract@open-review(Spotlight): Sliced-Wasserstein distance (SW) and its variant, Max Sliced-Wasserstein distance (Max-SW), have been used widely in the recent years due to their fast computation and scalability even when the probability measures lie in a very high dimensional space. However, SW requires many unnecessary projection samples to approximate its value while Max-SW only uses the most important projection, which ignores the information of other useful directions. In order to account for these weaknesses, we propose a novel distance, named Distributional Sliced-Wasserstein distance (DSW), that finds an optimal distribution over projections that can balance between exploring distinctive projecting directions and the informativeness of projections themselves. We show that the DSW is a generalization of Max-SW, and it can be computed efficiently by searching for the optimal push-forward measure over a set of probability measures over the unit sphere satisfying certain regularizing constraints that favor distinct directions. Finally, we conduct extensive experiments with large-scale datasets to demonstrate the favorable performances of the proposed distances over the previous sliced-based distances in generative modeling applications.

Evolving Reinforcement Learning Algorithms

- John D Co-Reyes, Yingjie Miao, Daiyi Peng, Esteban Real, Quoc V Le, Sergey Levine, Honglak Lee, Aleksandra Faust
- abstract@open-review(Oral): We propose a method for meta-learning reinforcement learning algorithms by searching over the space of computational graphs which compute the loss function for a value-based model-free RL agent to optimize. The learned algorithms are domain-agnostic and can generalize to new environments not seen during training. Our method can both learn from scratch and bootstrap off known existing algorithms, like DQN, enabling interpretable modifications which improve performance. Learning from scratch on simple classical control and gridworld tasks, our method rediscovers the temporal-difference (TD) algorithm. Bootstrapped from DQN, we highlight two learned algorithms which obtain good generalization performance over other classical control tasks, gridworld type tasks, and Atari games. The analysis of the learned algorithm behavior shows resemblance to recently proposed RL algorithms that address overestimation in value-based methods.

Systematic generalisation with group invariant predictions

- Faruk Ahmed, Yoshua Bengio, Harm van Seijen, Aaron Courville
- abstract@open-review(Spotlight): We consider situations where the presence of dominant simpler correlations with the target variable in a training set can cause an SGD-trained neural network to be less reliant on more persistently correlating complex features. When the non-persistent, simpler correlations correspond to non-semantic background factors, a neural network trained on this data can exhibit dramatic failure upon encountering systematic distributional shift, where the correlating background features are recombined with different objects. We perform an empirical study on three synthetic datasets, showing that group invariance methods across inferred partitionings of the training set can lead to significant improvements at such test-time situations. We also suggest a simple invariance penalty, showing with experiments on our setups that it can perform better than alternatives. We find that even without assuming access to any systematically shifted validation sets, one can still find improvements over an ERM-trained reference model.

Layer-adaptive Sparsity for the Magnitude-based Pruning

- Jaeho Lee, Sejun Park, Sangwoo Mo, Sungsoo Ahn, Jinwoo Shin
- abstract@open-review(Poster): Recent discoveries on neural network pruning reveal that, with a carefully chosen layerwise sparsity, a simple magnitude-based pruning achieves state-of-the-art tradeoff between sparsity and performance. However, without a clear consensus on ``how to choose,'' the layerwise sparsities are mostly selected algorithm-by-algorithm, often resorting to handcrafted heuristics or an extensive hyperparameter search. To fill this gap, we propose a novel importance score for global pruning, coined layer-adaptive magnitude-based pruning (LAMP) score; the score is a rescaled version of weight magnitude that incorporates the model-level ℓ_2 distortion incurred by pruning, and does not require any hyperparameter tuning or heavy computation. Under various image classification setups, LAMP consistently outperforms popular existing schemes for layerwise sparsity selection. Furthermore, we observe that LAMP continues to outperform baselines even in weight-rewinding setups, while the connectivity-oriented layerwise sparsity (the strongest baseline overall) performs worse than a simple global magnitude-based pruning in this case. Code: https://github.com/jaeho-lee/layer-adaptive-sparsity

Understanding and Improving Encoder Layer Fusion in Sequence-to-Sequence Learning

- Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, Zhaopeng Tu
- abstract@open-review(Poster): Encoder layer fusion (EncoderFusion) is a technique to fuse all the encoder layers (instead of the uppermost layer) for sequence-to-sequence (Seq2Seq) models, which has proven effective on various NLP tasks. However, it is still not entirely clear why and when EncoderFusion should work. In this paper, our main contribution is to take a step further in understanding EncoderFusion. Many of previous studies believe that the success of EncoderFusion comes from exploiting surface and syntactic information embedded in lower encoder layers. Unlike them, we find that the encoder embedding layer is more important than other intermediate encoder layers. In addition, the uppermost decoder layer consistently pays more attention to the encoder embedding layer across NLP tasks. Based on this observation, we propose a simple fusion method, SurfaceFusion, by fusing only the encoder embedding layer for the softmax layer. Experimental results show that SurfaceFusion outperforms EncoderFusion on several NLP benchmarks, including machine translation, text summarization, and grammatical error correction. It obtains the state-of-the-art performance on WMT16 Romanian-English and WMT14 English-French translation tasks. Extensive analyses reveal that SurfaceFusion learns more expressive bilingual word embeddings by building a closer relationship between relevant source and target embeddings. Source code is freely available at https://github.com/SunbowLiu/SurfaceFusion.

SaliencyMix: A Saliency Guided Data Augmentation Strategy for Better Regularization

- A F M Shahab Uddin, Mst. Sirazam Monira, Wheemyung Shin, TaeChoong Chung, Sung-Ho Bae
- abstract@open-review(Poster): Advanced data augmentation strategies have widely been studied to improve the generalization ability of deep learning models. Regional dropout is one of the popular solutions that guides the model to focus on less discriminative parts by randomly removing image regions, resulting in improved regularization. However, such information removal is undesirable. On the other hand, recent strategies suggest to randomly cut and mix patches and their labels among training images, to enjoy the advantages of regional dropout without having any pointless pixel in the augmented images. We argue that such random selection strategies of the patches may not necessarily represent sufficient information about the corresponding object and thereby mixing the labels according to that uninformative patch enables the model to learn unexpected feature representation. Therefore, we propose

SaliencyMix that carefully selects a representative image patch with the help of a saliency map and mixes this indicative patch with the target image, thus leading the model to learn more appropriate feature representation. SaliencyMix achieves the best known top-1 error of 21.26% and 20.09% for ResNet-50 and ResNet-101 architectures on ImageNet classification, respectively, and also improves the model robustness against adversarial perturbations. Furthermore, models that are trained with SaliencyMix, help to improve the object detection performance. Source code is available at \url{https://github.com/SaliencyMix/SaliencyMix}.

Image GANs meet Differentiable Rendering for Inverse Graphics and Interpretable 3D Neural Rendering

- Yuxuan Zhang, Wenzheng Chen, Huan Ling, Jun Gao, Yinan Zhang, Antonio Torralba, Sanja Fidler
- abstract@open-review(Oral): Differentiable rendering has paved the way to training neural networks to perform “inverse graphics” tasks such as predicting 3D geometry from monocular photographs. To train high performing models, most of the current approaches rely on multi-view imagery which are not readily available in practice. Recent Generative Adversarial Networks (GANs) that synthesize images, in contrast, seem to acquire 3D knowledge implicitly during training: object viewpoints can be manipulated by simply manipulating the latent codes. However, these latent codes often lack further physical interpretation and thus GANs cannot easily be inverted to perform explicit 3D reasoning. In this paper, we aim to extract and disentangle 3D knowledge learned by generative models by utilizing differentiable renderers. Key to our approach is to exploit GANs as a multi-view data generator to train an inverse graphics network using an off-the-shelf differentiable renderer, and the trained inverse graphics network as a teacher to disentangle the GAN's latent code into interpretable 3D properties. The entire architecture is trained iteratively using cycle consistency losses. We show that our approach significantly outperforms state-of-the-art inverse graphics networks trained on existing datasets, both quantitatively and via user studies. We further showcase the disentangled GAN as a controllable 3D “neural renderer", complementing traditional graphics renderers.

Are wider nets better given the same number of parameters?

- Anna Golubeva, Guy Gur-Ari, Behnam Neyshabur
- abstract@open-review(Poster): Empirical studies demonstrate that the performance of neural networks improves with increasing number of parameters. In most of these studies, the number of parameters is increased by increasing the network width. This begs the question: Is the observed improvement due to the larger number of parameters, or is it due to the larger width itself? We compare different ways of increasing model width while keeping the number of parameters constant. We show that for models initialized with a random, static sparsity pattern in the weight tensors, network width is the determining factor for good performance, while the number of weights is secondary, as long as the model achieves high training accuarcy. As a step towards understanding this effect, we analyze these models in the framework of Gaussian Process kernels. We find that the distance between the sparse finite-width model kernel and the infinite-width kernel at initialization is indicative of model performance.

Discovering Non-monotonic Autoregressive Orderings with Variational Inference

- Xuanlin Li, Brandon Trabucco, Dong Huk Park, Michael Luo, Sheng Shen, Trevor Darrell, Yang Gao
- abstract@open-review(Poster): The predominant approach for language modeling is to encode a sequence of tokens from left to right, but this eliminates a source of information: the order by which the sequence was naturally generated. One strategy to recover this information is to decode both the content and ordering of tokens. Some prior work supervises content and ordering with hand-designed loss functions to encourage specific orders or bootstraps from a predefined ordering. These approaches require domain-specific insight. Other prior work searches over valid insertion operations that lead to ground truth sequences during training, which has high time complexity and cannot be efficiently parallelized. We address these limitations with an unsupervised learner that can be trained in a fully-parallelizable manner to discover high-quality autoregressive orders in a data driven way without a domain-specific prior. The learner is a neural network that performs variational inference with the autoregressive ordering as a latent variable. Since the corresponding variational lower bound is not differentiable, we develop a practical algorithm for end-to-end optimization using policy gradients. Strong empirical results with our solution on sequence modeling tasks suggest that our algorithm is capable of discovering various autoregressive orders for different sequences that are competitive with or even better than fixed orders.

CompOFA – Compound Once-For-All Networks for Faster Multi-Platform Deployment

- Manas Sahni, Shreya Varshini, Alind Khare, Alexey Tumanov
- abstract@open-review(Poster): The emergence of CNNs in mainstream deployment has necessitated methods to design and train efficient architectures tailored to maximize the accuracy under diverse hardware and latency constraints. To scale these resource-intensive tasks with an increasing number of deployment targets, Once-For-All (OFA) proposed an approach to jointly train several models at once with a constant training cost. However, this cost remains as high as 40-50 GPU days and also suffers from a combinatorial explosion of sub-optimal model configurations. We seek to reduce this search space -- and hence the training budget -- by constraining search to models close to the accuracy-latency Pareto frontier. We incorporate insights of compound relationships between model dimensions to build CompOFA, a design space smaller by several orders of magnitude. Through experiments on ImageNet, we demonstrate that even with simple heuristics we can achieve a 2x reduction in training time and 216x speedup in model search/extraction time compared to the state of the art, without loss of Pareto optimality! We also show that this smaller design space is dense enough to support equally accurate models for a similar diversity of hardware and latency targets, while also reducing the complexity of the training and subsequent extraction algorithms. Our source code is available at https://github.com/gatech-sysml/CompOFA

Representing Partial Programs with Blended Abstract Semantics

- Maxwell Nye, Yewen Pu, Matthew Bowers, Jacob Andreas, Joshua B. Tenenbaum, Armando Solar-Lezama
- abstract@open-review(Poster): Synthesizing programs from examples requires searching over a vast, combinatorial space of possible programs. In this search process, a key challenge is representing the behavior of a partially written program before it can be executed, to judge if it is on the right track and predict where to search next. We introduce a general technique for representing partially written programs in a program synthesis engine. We take inspiration from the technique of abstract interpretation, in which an approximate execution model is used to determine if an unfinished program will eventually satisfy a goal specification. Here we learn an approximate execution model implemented as a modular neural network. By constructing compositional program representations that implicitly encode the interpretation semantics of the underlying programming language, we can represent partial programs using a flexible combination of concrete execution state and learned neural representations, using the learned approximate semantics when concrete semantics are not known (in unfinished parts of the program). We show that these hybrid neuro-symbolic representations enable execution-guided synthesizers to use more powerful language constructs, such as loops and higher-order functions, and can be used to synthesize programs more accurately for a given search budget than pure neural approaches in several domains.

MONGOOSE: A Learnable LSH Framework for Efficient Neural Network Training

- Beidi Chen, Zichang Liu, Binghui Peng, Zhaozhuo Xu, Jonathan Lingjie Li, Tri Dao, Zhao Song, Anshumali Shrivastava, Christopher Re
- abstract@open-review(Oral): Recent advances by practitioners in the deep learning community have breathed new life into Locality Sensitive Hashing (LSH), using it to reduce memory and time bottlenecks in neural network (NN) training. However, while LSH has sub-linear guarantees for approximate near-neighbor search in theory, it is known to have inefficient query time in practice due to its use of random hash functions. Moreover, when model parameters are changing, LSH suffers from update overhead. This work is motivated by an observation that model parameters evolve slowly, such that the changes do not always require an LSH update to maintain performance. This phenomenon points to the potential for a reduction in update time and allows for a modified learnable version of data-dependent LSH to improve query time at a low cost. We use the above insights to build MONGOOSE, an end-to-

end LSH framework for efficient NN training. In particular, MONGOOSE is equipped with a scheduling algorithm to adaptively perform LSH updates with provable guarantees and learnable hash functions to improve query efficiency. Empirically, we validate MONGOOSE on large-scale deep learning models for recommendation systems and language modeling. We find that it achieves up to 8% better accuracy compared to previous LSH approaches, with $6.5 \times$ speed-up and $6\times$ reduction in memory usage.

PolarNet: Learning to Optimize Polar Keypoints for Keypoint Based Object Detection

- Wu Xiongwei, Doyen Sahoo, Steven HOI
- abstract@open-review(Poster): A variety of anchor-free object detectors have been actively proposed as possible alternatives to the mainstream anchor-based detectors that often rely on complicated design of anchor boxes. Despite achieving promising performance on par with anchor-based detectors, the existing anchor-free detectors such as FCOS or CenterNet predict objects based on standard Cartesian coordinates, which often yield poor quality keypoints. Further, the feature representation is also scale-sensitive. In this paper, we propose a new anchor-free keypoint based detector `PolarNet", where keypoints are represented as a set of Polar coordinates instead of Cartesian coordinates. The`PolarNet" detector learns offsets pointing to the corners of objects in order to learn high quality keypoints. Additionally, PolarNet uses features of corner points to localize objects, making the localization scale-insensitive. Finally in our experiments, we show that PolarNet, an anchor-free detector, outperforms the existing anchor-free detectors, and it is able to achieve highly competitive result on COCO test-dev benchmark (47.8% and 50.3% AP under the single-model single-scale and multi-scale testing) which is on par with the state-of-the-art two-stage anchor-based object detectors. The code and the models are available at https://github.com/XiongweiWu/PolarNetV1

Neural Architecture Search on ImageNet in Four GPU Hours: A Theoretically Inspired Perspective

- Wuyang Chen, Xinyu Gong, Zhangyang Wang
- abstract@open-review(Poster): Neural Architecture Search (NAS) has been explosively studied to automate the discovery of top-performer neural networks. Current works require heavy training of supernet or intensive architecture evaluations, thus suffering from heavy resource consumption and often incurring search bias due to truncated training or approximations. Can we select the best neural architectures without involving any training and eliminate a drastic portion of the search cost?

We provide an affirmative answer, by proposing a novel framework called \textit{training-free neural architecture search} ($\textbf{TE-NAS}$). TE-NAS ranks architectures by analyzing the spectrum of the neural tangent kernel (NTK), and the number of linear regions in the input space. Both are motivated by recent theory advances in deep networks, and can be computed without any training. We show that: (1) these two measurements imply the $\textit{trainability}$ and $\textit{expressivity}$ of a neural network; and (2) they strongly correlate with the network's actual test accuracy. Further on, we design a pruning-based NAS mechanism to achieve a more flexible and superior trade-off between the trainability and expressivity during the search. In NAS-Bench-201 and DARTS search spaces, TE-NAS completes high-quality search but only costs $\textbf{0.5}$ and $\textbf{4}$ GPU hours with one 1080Ti on CIFAR-10 and ImageNet, respectively. We hope our work to inspire more attempts in bridging between the theoretic findings of deep networks and practical impacts in real NAS applications.

Unsupervised Representation Learning for Time Series with Temporal Neighborhood Coding

- Sana Tonekaboni, Danny Eytan, Anna Goldenberg
- abstract@open-review(Poster): Time series are often complex and rich in information but sparsely labeled and therefore challenging to model. In this paper, we propose a self-supervised framework for learning robust and generalizable representations for time series. Our approach, called Temporal Neighborhood Coding (TNC), takes advantage of the local smoothness of a signal's generative process to define neighborhoods in time with stationary properties. Using a debiased contrastive objective, our framework learns time series representations by ensuring that in the encoding space, the distribution of signals from within a neighborhood is distinguishable from the distribution of non-neighboring signals. Our motivation stems from the medical field, where the ability to model the dynamic nature of time series data is especially valuable for identifying, tracking, and predicting the underlying patients' latent states in settings where labeling data is practically impossible. We compare our method to recently developed unsupervised representation learning approaches and demonstrate superior performance on clustering and classification tasks for multiple datasets.

Do Wide and Deep Networks Learn the Same Things? Uncovering How Neural Network Representations Vary with Width and Depth

- Thao Nguyen, Maithra Raghu, Simon Kornblith
- abstract@open-review(Poster): A key factor in the success of deep neural networks is the ability to scale models to improve performance by varying the architecture depth and width. This simple property of neural network design has resulted in highly effective architectures for a variety of tasks. Nevertheless, there is limited understanding of effects of depth and width on the learned representations. In this paper, we study this fundamental question. We begin by investigating how varying depth and width affects model hidden representations, finding a characteristic block structure in the hidden representations of larger capacity (wider or deeper) models. We demonstrate that this block structure arises when model capacity is large relative to the size of the training set, and is indicative of the underlying layers preserving and propagating the dominant principal component of their representations. This discovery has important ramifications for features learned by different models, namely, representations outside the block structure are often similar across architectures with varying widths and depths, but the block structure is unique to each model. We analyze the output predictions of different model architectures, finding that even when the overall accuracy is similar, wide and deep models exhibit distinctive error patterns and variations across classes.

Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning

- Beliz Gunel, Jingfei Du, Alexis Conneau, Veselin Stoyanov
- abstract@open-review(Poster): State-of-the-art natural language understanding classification models follow two-stages: pre-training a large language model on an auxiliary task, and then fine-tuning the model on a task-specific labeled dataset using cross-entropy loss. However, the cross-entropy loss has several shortcomings that can lead to sub-optimal generalization and instability. Driven by the intuition that good generalization requires capturing the similarity between examples in one class and contrasting them with examples in other classes, we propose a supervised contrastive learning (SCL) objective for the fine-tuning stage. Combined with cross-entropy, our proposed SCL loss obtains significant improvements over a strong RoBERTa-Large baseline on multiple datasets of the GLUE benchmark in few-shot learning settings, without requiring specialized architecture, data augmentations, memory banks, or additional unsupervised data. Our proposed fine-tuning objective leads to models that are more robust to different levels of noise in the fine-tuning training data, and can generalize better to related tasks with limited labeled data.

Memory Optimization for Deep Networks

- Aashaka Shah, Chao-Yuan Wu, Jayashree Mohan, Vijay Chidambaram, Philipp Kraehenbuehl
- abstract@open-review(Spotlight): Deep learning is slowly, but steadily, hitting a memory bottleneck. While the tensor computation in top-of-the-line GPUs increased by $32\times$ over the last five years, the total available memory only grew by $2.5\times$. This prevents researchers from exploring larger architectures, as training large networks requires more memory for storing intermediate outputs. In this paper, we present MONeT, an automatic framework that minimizes both the memory footprint and computational overhead of deep networks. MONeT jointly optimizes the checkpointing schedule and the implementation of various operators. MONeT is able to outperform all prior hand-tuned operations as well as automated checkpointing.

MONeT reduces the overall memory requirement by $3\times$ for various PyTorch models, with a 9-16$\%$ overhead in computation. For the same computation cost, MONeT requires 1.2-1.8$\times$ less memory than current state-of-the-art automated checkpointing frameworks. Our code will be made publicly available upon acceptance.

Parrot: Data-Driven Behavioral Priors for Reinforcement Learning

- Avi Singh, Huihan Liu, Gaoyue Zhou, Albert Yu, Nicholas Rhinehart, Sergey Levine
- abstract@open-review(Oral): Reinforcement learning provides a general framework for flexible decision making and control, but requires extensive data collection for each new task that an agent needs to learn. In other machine learning fields, such as natural language processing or computer vision, pre-training on large, previously collected datasets to bootstrap learning for new tasks has emerged as a powerful paradigm to reduce data requirements when learning a new task. In this paper, we ask the following question: how can we enable similarly useful pre-training for RL agents? We propose a method for pre-training behavioral priors that can capture complex input-output relationships observed in successful trials from a wide range of previously seen tasks, and we show how this learned prior can be used for rapidly learning new tasks without impeding the RL agent's ability to try out novel behaviors. We demonstrate the effectiveness of our approach in challenging robotic manipulation domains involving image observations and sparse reward functions, where our method outperforms prior works by a substantial margin. Additional materials can be found on our project website: https://sites.google.com/view/parrot-rl

Early Stopping in Deep Networks: Double Descent and How to Eliminate it

- Reinhard Heckel, Fatih Furkan Yilmaz
- abstract@open-review(Poster): Over-parameterized models, such as large deep networks, often exhibit a double descent phenomenon, whereas a function of model size, error first decreases, increases, and decreases at last. This intriguing double descent behavior also occurs as a function of training epochs and has been conjectured to arise because training epochs control the model complexity. In this paper, we show that such epoch-wise double descent occurs for a different reason: It is caused by a superposition of two or more bias-variance tradeoffs that arise because different parts of the network are learned at different epochs, and mitigating this by proper scaling of stepsizes can significantly improve the early stopping performance. We show this analytically for i) linear regression, where differently scaled features give rise to a superposition of bias-variance tradeoffs, and for ii) a wide two-layer neural network, where the first and second layers govern bias-variance tradeoffs. Inspired by this theory, we study two standard convolutional networks empirically and show that eliminating epoch-wise double descent through adjusting stepsizes of different layers improves the early stopping performance.

Contrastive Syn-to-Real Generalization

- Wuyang Chen, Zhiding Yu, Shalini De Mello, Sifei Liu, Jose M. Alvarez, Zhangyang Wang, Anima Anandkumar
- abstract@open-review(Poster): Training on synthetic data can be beneficial for label or data-scarce scenarios. However, synthetically trained models often suffer from poor generalization in real domains due to domain gaps. In this work, we make a key observation that the diversity of the learned feature embeddings plays an important role in the generalization performance. To this end, we propose contrastive synthetic-to-real generalization (CSG), a novel framework that leverage the pre-trained ImageNet knowledge to prevent overfitting to the synthetic domain, while promoting the diversity of feature embeddings as an inductive bias to improve generalization. In addition, we enhance the proposed CSG framework with attentional pooling (A-pool) to let the model focus on semantically important regions and further improve its generalization. We demonstrate the effectiveness of CSG on various synthetic training tasks, exhibiting state-of-the-art performance on zero-shot domain generalization.

Benchmarks for Deep Off-Policy Evaluation

- Justin Fu, Mohammad Norouzi, Ofir Nachum, George Tucker, ziyu wang, Alexander Novikov, Mengjiao Yang, Michael R Zhang, Yutian Chen, Aviral Kumar, Cosmin Paduraru, Sergey Levine, Thomas Paine
- abstract@open-review(Poster): Off-policy evaluation (OPE) holds the promise of being able to leverage large, offline datasets for both evaluating and selecting complex policies for decision making. The ability to learn offline is particularly important in many real-world domains, such as in healthcare, recommender systems, or robotics, where online data collection is an expensive and potentially dangerous process. Being able to accurately evaluate and select high-performing policies without requiring online interaction could yield significant benefits in safety, time, and cost for these applications. While many OPE methods have been proposed in recent years, comparing results between papers is difficult because currently there is a lack of a comprehensive and unified benchmark, and measuring algorithmic progress has been challenging due to the lack of difficult evaluation tasks. In order to address this gap, we present a collection of policies that in conjunction with existing offline datasets can be used for benchmarking off-policy evaluation. Our tasks include a range of challenging high-dimensional continuous control problems, with wide selections of datasets and policies for performing policy selection. The goal of our benchmark is to provide a standardized measure of progress that is motivated from a set of principles designed to challenge and test the limits of existing OPE methods. We perform an evaluation of state-of-the-art algorithms and provide open-source access to our data and code to foster future research in this area.

Pre-training Text-to-Text Transformers for Concept-centric Common Sense

- Wangchunshu Zhou, Dong-Ho Lee, Ravi Kiran Selvam, Seyeon Lee, Xiang Ren
- abstract@open-review(Poster): Pretrained language models (PTLM) have achieved impressive results in a range of natural language understanding (NLU) and generation (NLG) tasks that require a syntactic and semantic understanding of the text. However, current pre-training objectives such as masked token prediction (for BERT-style PTLMs) and masked span infilling (for T5-style PTLMs) do not explicitly model the relational and compositional commonsense knowledge about everyday concepts, which is crucial to many downstream tasks requiring commonsense reasoning. To augment PTLMs with common sense, we propose generative and contrastive objectives as intermediate self-supervised pre-training tasks between general pre-training and downstream task-specific fine-tuning. We also propose a joint training framework to unify generative and contrastive objectives so that these objectives can be more effective. Our proposed objectives can pack more commonsense knowledge into the parameters of a pre-trained text-to-text transformer without relying on external knowledge bases, yielding better performance on both NLU and NLG tasks. We apply our method on a pre-trained T5 model in an intermediate task transfer learning fashion to train a concept-aware language model (CALM) and experiment with five commonsense benchmarks (four NLU tasks and one NLG task). Experimental results show that CALM outperforms baseline methods by a consistent margin.

Combining Label Propagation and Simple Models out-performs Graph Neural Networks

- Qian Huang, Horace He, Abhay Singh, Ser-Nam Lim, Austin Benson
- abstract@open-review(Poster): Graph Neural Networks (GNNs) are a predominant technique for learning over graphs. However, there is relatively little understanding of why GNNs are successful in practice and whether they are necessary for good performance. Here, we show that for many standard transductive node classification benchmarks, we can exceed or match the performance of state-of-the-art GNNs by combining shallow models that ignore the graph structure with two simple post-processing steps that exploit correlation in the label structure: (i) an “error correlation” that spreads residual errors in training data to correct errors in test data and (ii) a “prediction correlation” that smooths the predictions on the test data. We call this overall procedure Correct and Smooth (C&S), and the post-processing steps are implemented via simple modifications to standard label propagation techniques that have long been used in graph-based semi-supervised learning. Our approach exceeds or nearly matches the performance of state-of-the-art GNNs on a wide variety of benchmarks, with just a small fraction of the parameters and orders of magnitude faster runtime. For instance, we exceed the best-known GNN performance on the OGB-Products dataset with 137 times fewer parameters and greater than 100 times less training time. The performance of our

methods highlights how directly incorporating label information into the learning algorithm (as is common in traditional methods) yields easy and substantial performance gains. We can also incorporate our techniques into big GNN models, providing modest gains in some cases.

Learning Long-term Visual Dynamics with Region Proposal Interaction Networks

- Haozhi Qi, Xiaolong Wang, Deepak Pathak, Yi Ma, Jitendra Malik
- abstract@open-review(Poster): Learning long-term dynamics models is the key to understanding physical common sense. Most existing approaches on learning dynamics from visual input sidestep long-term predictions by resorting to rapid re-planning with short-term models. This not only requires such models to be super accurate but also limits them only to tasks where an agent can continuously obtain feedback and take action at each step until completion. In this paper, we aim to leverage the ideas from success stories in visual recognition tasks to build object representations that can capture inter-object and object-environment interactions over a long range. To this end, we propose Region Proposal Interaction Networks (RPIN), which reason about each object's trajectory in a latent region-proposal feature space. Thanks to the simple yet effective object representation, our approach outperforms prior methods by a significant margin both in terms of prediction quality and their ability to plan for downstream tasks, and also generalize well to novel environments. Code, pre-trained models, and more visualization results are available at https://haozhi.io/RPIN.

Chaos of Learning Beyond Zero-sum and Coordination via Game Decompositions

- Yun Kuen Cheung, Yixin Tao
- abstract@open-review(Poster): It is of primary interest for ML to understand how agents learn and interact dynamically in competitive environments and games (e.g. GANs). But this has been a difficult task, as irregular behaviors are commonly observed in such systems. This can be explained theoretically, for instance, by the works of Cheung and Piliouras (COLT 2019; NeurIPS 2020), which showed that in two-person zero-sum games, if agents employ one of the most well-known learning algorithms, Multiplicative Weights Update (MWU), then Lyapunov chaos occurs everywhere in the payoff space. In this paper, we study how persistent chaos can occur in the more general normal game settings, where the agents might have the motivation to coordinate (which is not true for zero-sum games) and the number of agents can be arbitrary.

We characterize bimatrix games where MWU, its optimistic variant (OMWU) or Follow-the-Regularized-Leader (FTRL) algorithms are Lyapunov chaotic almost everywhere in the payoff space. Technically, our characterization is derived by extending the volume-expansion argument of Cheung and Piliouras via the canonical game decomposition into zero-sum and coordination components. Interestingly, the two components induce opposite volume-changing behaviors, so the overall behavior can be analyzed by comparing the strengths of the components against each other. The comparison is done via our new notion of "matrix domination" or via a linear program. For multi-player games, we present a local equivalence of volume change between general games and graphical games, which is used to perform volume and chaos analyses of MWU and OMWU in potential games.

Control-Aware Representations for Model-based Reinforcement Learning

- Brandon Cui, Yinlam Chow, Mohammad Ghavamzadeh
- abstract@open-review(Poster): A major challenge in modern reinforcement learning (RL) is efficient control of dynamical systems from high-dimensional sensory observations. Learning controllable embedding (LCE) is a promising approach that addresses this challenge by embedding the observations into a lower-dimensional latent space, estimating the latent dynamics, and utilizing it to perform control in the latent space. Two important questions in this area are how to learn a representation that is amenable to the control problem at hand, and how to achieve an end-to-end framework for representation learning and control. In this paper, we take a few steps towards addressing these questions. We first formulate a LCE model to learn representations that are suitable to be used by a policy iteration style algorithm in the latent space.We call this model control-aware representation learning(CARL). We derive a loss function and three implementations for CARL. In the offline implementation, we replace the locally-linear control algorithm (e.g., iLQR) used by the existing LCE methods with a RL algorithm, namely model-based soft actor-critic, and show that it results in significant improvement. In online CARL, we interleave representation learning and control, and demonstrate further gain in performance. Finally, we propose value-guided CARL, a variation in which we optimize a weighted version of the CARL loss function, where the weights depend on the TD-error of the current policy. We evaluate the proposed algorithms by extensive experiments on benchmark tasks and compare them with several LCE baselines.

Provably robust classification of adversarial examples with detection

- Fatemeh Sheikholeslami, Ali Lotfi, J Zico Kolter
- abstract@open-review(Poster): Adversarial attacks against deep networks can be defended against either by building robust classifiers or, by creating classifiers that can \emph{detect} the presence of adversarial perturbations. Although it may intuitively seem easier to simply detect attacks rather than build a robust classifier, this has not bourne out in practice even empirically, as most detection methods have subsequently been broken by adaptive attacks, thus necessitating \emph{verifiable} performance for detection mechanisms. In this paper, we propose a new method for jointly training a provably robust classifier and detector. Specifically, we show that by introducing an additional "abstain/detection" into a classifier, we can modify existing certified defense mechanisms to allow the classifier to either robustly classify \emph{or} detect adversarial attacks. We extend the common interval bound propagation (IBP) method for certified robustness under ℓ_∞ perturbations to account for our new robust objective, and show that the method outperforms traditional IBP used in isolation, especially for large perturbation sizes. Specifically, tests on MNIST and CIFAR-10 datasets exhibit promising results, for example with provable robust error less than 63.63% and 67.92%, for 55.6% and 66.37% natural error, for $\epsilon=8/255$ and $16/255$ on the CIFAR-10 dataset, respectively.

Return-Based Contrastive Representation Learning for Reinforcement Learning

- Guoqing Liu, Chuheng Zhang, Li Zhao, Tao Qin, Jinhua Zhu, Li Jian, Nenghai Yu, Tie-Yan Liu
- abstract@open-review(Poster): Recently, various auxiliary tasks have been proposed to accelerate representation learning and improve sample efficiency in deep reinforcement learning (RL). However, existing auxiliary tasks do not take the characteristics of RL problems into consideration and are unsupervised. By leveraging returns, the most important feedback signals in RL, we propose a novel auxiliary task that forces the learnt representations to discriminate state-action pairs with different returns. Our auxiliary loss is theoretically justified to learn representations that capture the structure of a new form of state-action abstraction, under which state-action pairs with similar return distributions are aggregated together. Empirically, our algorithm outperforms strong baselines on complex tasks in Atari games and DeepMind Control suite, and achieves even better performance when combined with existing auxiliary tasks.

Adversarially-Trained Deep Nets Transfer Better: Illustration on Image Classification

- Francisco Utrera, Evan Kravitz, N. Benjamin Erichson, Rajiv Khanna, Michael W. Mahoney
- abstract@open-review(Poster): Transfer learning has emerged as a powerful methodology for adapting pre-trained deep neural networks on image recognition tasks to new domains. This process consists of taking a neural network pre-trained on a large feature-rich source dataset, freezing the early layers that encode essential generic image properties, and then fine-tuning the last few layers in order to capture specific information related to the target situation. This approach is particularly useful when only limited or weakly labeled data are available for the new task. In this work, we demonstrate that adversarially-trained models transfer better than non-adversarially-trained models, especially if only limited data are available for the new domain task. Further, we observe that adversarial training biases the learnt representations to retaining shapes, as opposed to textures, which impacts the transferability

of the source models. Finally, through the lens of influence functions, we discover that transferred adversarially-trained models contain more human-identifiable semantic information, which explains -- at least partly -- why adversarially-trained models transfer better.

Learning Structural Edits via Incremental Tree Transformations

- Ziyu Yao, Frank F. Xu, Pengcheng Yin, Huan Sun, Graham Neubig
- abstract@open-review(Poster): While most neural generative models generate outputs in a single pass, the human creative process is usually one of iterative building and refinement. Recent work has proposed models of editing processes, but these mostly focus on editing sequential data and/or only model a single editing pass. In this paper, we present a generic model for incremental editing of structured data (i.e. "structural edits"). Particularly, we focus on tree-structured data, taking abstract syntax trees of computer programs as our canonical example. Our editor learns to iteratively generate tree edits (e.g. deleting or adding a subtree) and applies them to the partially edited data, thereby the entire editing process can be formulated as consecutive, incremental tree transformations. To show the unique benefits of modeling tree edits directly, we further propose a novel edit encoder for learning to represent edits, as well as an imitation learning method that allows the editor to be more robust. We evaluate our proposed editor on two source code edit datasets, where results show that, with the proposed edit encoder, our editor significantly improves accuracy over previous approaches that generate the edited program directly in one pass. Finally, we demonstrate that training our editor to imitate experts and correct its mistakes dynamically can further improve its performance.

Cross-Attentional Audio-Visual Fusion for Weakly-Supervised Action Localization

- Jun-Tae Lee, Mihir Jain, Hyoungwoo Park, Sungrack Yun
- abstract@open-review(Poster): Temporally localizing actions in videos is one of the key components for video understanding. Learning from weakly-labeled data is seen as a potential solution towards avoiding expensive frame-level annotations. Different from other works which only depend on visual-modality, we propose to learn richer audiovisual representation for weakly-supervised action localization. First, we propose a multi-stage cross-attention mechanism to collaboratively fuse audio and visual features, which preserves the intra-modal characteristics. Second, to model both foreground and background frames, we construct an open-max classifier that treats the background class as an open-set. Third, for precise action localization, we design consistency losses to enforce temporal continuity for the action class prediction, and also help with foreground-prediction reliability. Extensive experiments on two publicly available video-datasets (AVE and ActivityNet1.2) show that the proposed method effectively fuses audio and visual modalities, and achieves the state-of-the-art results for weakly-supervised action localization.

Improved Estimation of Concentration Under ℓ_p-Norm Distance Metrics Using Half Spaces

- Jack Prescott, Xiao Zhang, David Evans
- abstract@open-review(Poster): Concentration of measure has been argued to be the fundamental cause of adversarial vulnerability. Mahloujifar et al. (2019) presented an empirical way to measure the concentration of a data distribution using samples, and employed it to find lower bounds on intrinsic robustness for several benchmark datasets. However, it remains unclear whether these lower bounds are tight enough to provide a useful approximation for the intrinsic robustness of a dataset. To gain a deeper understanding of the concentration of measure phenomenon, we first extend the Gaussian Isoperimetric Inequality to non-spherical Gaussian measures and arbitrary ℓ_p-norms ($p \geq 2$). We leverage these theoretical insights to design a method that uses half-spaces to estimate the concentration of any empirical dataset under ℓ_p-norm distance metrics. Our proposed algorithm is more efficient than Mahloujifar et al. (2019)'s, and experiments on synthetic datasets and image benchmarks demonstrate that it is able to find much tighter intrinsic robustness bounds. These tighter estimates provide further evidence that rules out intrinsic dataset concentration as a possible explanation for the adversarial vulnerability of state-of-the-art classifiers.

Beyond Categorical Label Representations for Image Classification

- Boyuan Chen, Yu Li, Sunand Raghupathi, Hod Lipson
- abstract@open-review(Poster): We find that the way we choose to represent data labels can have a profound effect on the quality of trained models. For example, training an image classifier to regress audio labels rather than traditional categorical probabilities produces a more reliable classification. This result is surprising, considering that audio labels are more complex than simpler numerical probabilities or text. We hypothesize that high dimensional, high entropy label representations are generally more useful because they provide a stronger error signal. We support this hypothesis with evidence from various label representations including constant matrices, spectrograms, shuffled spectrograms, Gaussian mixtures, and uniform random matrices of various dimensionalities. Our experiments reveal that high dimensional, high entropy labels achieve comparable accuracy to text (categorical) labels on standard image classification tasks, but features learned through our label representations exhibit more robustness under various adversarial attacks and better effectiveness with a limited amount of training data. These results suggest that label representation may play a more important role than previously thought.

Fantastic Four: Differentiable and Efficient Bounds on Singular Values of Convolution Layers

- Sahil Singla, Soheil Feizi
- abstract@open-review(Poster): In deep neural networks, the spectral norm of the Jacobian of a layer bounds the factor by which the norm of a signal changes during forward/backward propagation. Spectral norm regularizations have been shown to improve generalization, robustness and optimization of deep learning methods. Existing methods to compute the spectral norm of convolution layers either rely on heuristics that are efficient in computation but lack guarantees or are theoretically-sound but computationally expensive. In this work, we obtain the best of both worlds by deriving {\it four} provable upper bounds on the spectral norm of a standard 2D multi-channel convolution layer. These bounds are differentiable and can be computed efficiently during training with negligible overhead. One of these bounds is in fact the popular heuristic method of Miyato et al. (multiplied by a constant factor depending on filter sizes). Each of these four bounds can achieve the tightest gap depending on convolution filters. Thus, we propose to use the minimum of these four bounds as a tight, differentiable and efficient upper bound on the spectral norm of convolution layers. Moreover, our spectral bound is an effective regularizer and can be used to bound either the lipschitz constant or curvature values (eigenvalues of the Hessian) of neural networks. Through experiments on MNIST and CIFAR-10, we demonstrate the effectiveness of our spectral bound in improving generalization and robustness of deep networks.

Accelerating Convergence of Replica Exchange Stochastic Gradient MCMC via Variance Reduction

- Wei Deng, Qi Feng, Georgios P. Karagiannis, Guang Lin, Faming Liang
- abstract@open-review(Poster): Replica exchange stochastic gradient Langevin dynamics (reSGLD) has shown promise in accelerating the convergence in non-convex learning; however, an excessively large correction for avoiding biases from noisy energy estimators has limited the potential of the acceleration. To address this issue, we study the variance reduction for noisy energy estimators, which promotes much more effective swaps. Theoretically, we provide a non-asymptotic analysis on the exponential convergence for the underlying continuous-time Markov jump process; moreover, we consider a generalized Girsanov theorem which includes the change of Poisson measure to overcome the crude discretization based on the Gr\"{o}wall's inequality and yields a much tighter error in the 2-Wasserstein ($\mathcal{W}_2$) distance. Numerically, we conduct extensive experiments and obtain state-of-the-art results in optimization and uncertainty estimates for synthetic experiments and image data.

IsarStep: a Benchmark for High-level Mathematical Reasoning

- Wenda Li, Lei Yu, Yuhuai Wu, Lawrence C. Paulson
- abstract@open-review(Poster): A well-defined benchmark is essential for measuring and accelerating research progress of machine learning models. In this paper, we present a benchmark for high-level mathematical reasoning and study the reasoning capabilities of neural sequence-to-sequence models. We build a non-synthetic dataset from the largest repository of proofs written by human experts in a theorem prover. The dataset has a broad coverage of undergraduate and research-level mathematical and computer science theorems. In our defined task, a model is required to fill in a missing intermediate proposition given surrounding proofs. This task provides a starting point for the long-term goal of having machines generate human-readable proofs automatically. Our experiments and analysis reveal that while the task is challenging, neural models can capture non-trivial mathematical reasoning. We further design a hierarchical transformer that outperforms the transformer baseline.

HW-NAS-Bench: Hardware-Aware Neural Architecture Search Benchmark

- Chaojian Li, Zhongzhi Yu, Yonggan Fu, Yongan Zhang, Yang Zhao, Haoran You, Qixuan Yu, Yue Wang, Cong Hao, Yingyan Lin
- abstract@open-review(Spotlight): HardWare-aware Neural Architecture Search (HW-NAS) has recently gained tremendous attention by automating the design of deep neural networks deployed in more resource-constrained daily life devices. Despite its promising performance, developing optimal HW-NAS solutions can be prohibitively challenging as it requires cross-disciplinary knowledge in the algorithm, micro-architecture, and device-specific compilation. First, to determine the hardware-cost to be incorporated into the NAS process, existing works mostly adopt either pre-collected hardware-cost look-up tables or device-specific hardware-cost models. The former can be time-consuming due to the required knowledge of the device's compilation method and how to set up the measurement pipeline, while building the latter is often a barrier for non-hardware experts like NAS researchers. Both of them limit the development of HW-NAS innovations and impose a barrier-to-entry to non-hardware experts. Second, similar to generic NAS, it can be notoriously difficult to benchmark HW-NAS algorithms due to their significant required computational resources and the differences in adopted search spaces, hyperparameters, and hardware devices. To this end, we develop HW-NAS-Bench, the first public dataset for HW-NAS research which aims to democratize HW-NAS research to non-hardware experts and make HW-NAS research more reproducible and accessible. To design HW-NAS-Bench, we carefully collected the measured/estimated hardware performance (e.g., energy cost and latency) of all the networks in the search spaces of both NAS-Bench-201 and FBNet, on six hardware devices that fall into three categories (i.e., commercial edge devices, FPGA, and ASIC). Furthermore, we provide a comprehensive analysis of the collected measurements in HW-NAS-Bench to provide insights for HW-NAS research. Finally, we demonstrate exemplary user cases to (1) show that HW-NAS-Bench allows non-hardware experts to perform HW-NAS by simply querying our pre-measured dataset and (2) verify that dedicated device-specific HW-NAS can indeed lead to optimal accuracy-cost trade-offs. The codes and all collected data are available at https://github.com/RICE-EIC/HW-NAS-Bench.

Factorizing Declarative and Procedural Knowledge in Structured, Dynamical Environments

- Anirudh Goyal, Alex Lamb, Phanideep Gampa, Philippe Beaudoin, Charles Blundell, Sergey Levine, Yoshua Bengio, Michael Curtis Mozer
- abstract@open-review(Poster): Modeling a structured, dynamic environment like a video game requires keeping track of the objects and their states (declarative knowledge) as well as predicting how objects behave (procedural knowledge). Black-box models with a monolithic hidden state often fail to apply procedural knowledge consistently and uniformly, i.e., they lack systematicity. For example, in a video game, correct prediction of one enemy's trajectory does not ensure correct prediction of another's. We address this issue via an architecture that factorizes declarative and procedural knowledge and that imposes modularity within each form of knowledge. The architecture consists of active modules called object files that maintain the state of a single object and invoke passive external knowledge sources called schemata that prescribe state updates. To use a video game as an illustration, two enemies of the same type will share schemata but will have separate object files to encode their distinct state (e.g., health, position). We propose to use attention to determine which object files to update, the selection of schemata, and the propagation of information between object files. The resulting architecture is a drop-in replacement conforming to the same input-output interface as normal recurrent networks (e.g., LSTM, GRU) yet achieves substantially better generalization on environments that have multiple object tokens of the same type, including a challenging intuitive physics benchmark.

Provable Rich Observation Reinforcement Learning with Combinatorial Latent States

- Dipendra Misra, Qinghua Liu, Chi Jin, John Langford
- abstract@open-review(Poster): We propose a novel setting for reinforcement learning that combines two common real-world difficulties: presence of observations (such as camera images) and factored states (such as location of objects). In our setting, the agent receives observations generated stochastically from a "latent" factored state. These observations are "rich enough" to enable decoding of the latent state and remove partial observability concerns. Since the latent state is combinatorial, the size of state space is exponential in the number of latent factors. We create a learning algorithm FactoRL (Fact-o-Rel) for this setting, which uses noise-contrastive learning to identify latent structures in emission processes and discover a factorized state space. We derive polynomial sample complexity guarantees for FactoRL which polynomially depend upon the number factors, and very weakly depend on the size of the observation space. We also provide a guarantee of polynomial time complexity when given access to an efficient planning algorithm.

LowKey: Leveraging Adversarial Attacks to Protect Social Media Users from Facial Recognition

- Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John P Dickerson, Gavin Taylor, Tom Goldstein
- abstract@open-review(Poster): Facial recognition systems are increasingly deployed by private corporations, government agencies, and contractors for consumer services and mass surveillance programs alike. These systems are typically built by scraping social media profiles for user images. Adversarial perturbations have been proposed for bypassing facial recognition systems. However, existing methods fail on full-scale systems and commercial APIs. We develop our own adversarial filter that accounts for the entire image processing pipeline and is demonstrably effective against industrial-grade pipelines that include face detection and large scale databases. Additionally, we release an easy-to-use webtool that significantly degrades the accuracy of Amazon Rekognition and the Microsoft Azure Face Recognition API, reducing the accuracy of each to below 1%.

Neural Networks for Learning Counterfactual G-Invariances from Single Environments

- S Chandra Mouli, Bruno Ribeiro
- abstract@open-review(Poster): Despite —or maybe because of— their astonishing capacity to fit data, neural networks are believed to have difficulties extrapolating beyond training data distribution. This work shows that, for extrapolations based on finite transformation groups, a model's inability to extrapolate is unrelated to its capacity. Rather, the shortcoming is inherited from a learning hypothesis: Examples not explicitly observed with infinitely many training examples have underspecified outcomes in the learner's model. In order to endow neural networks with the ability to extrapolate over group transformations, we introduce a learning framework counterfactually-guided by the learning hypothesis that any group invariance to (known) transformation groups is mandatory even without evidence, unless the learner deems it inconsistent with the training data. Unlike existing invariance-driven methods for (counterfactual) extrapolations, this framework allows extrapolations from a single environment. Finally, we introduce sequence and image extrapolation tasks that validate our framework and showcase the shortcomings of traditional approaches.

Simple Spectral Graph Convolution

- Hao Zhu, Piotr Koniusz
- abstract@open-review(Poster): Graph Convolutional Networks (GCNs) are leading methods for learning graph representations. However, without specially designed architectures, the performance of GCNs degrades quickly with increased depth. As the aggregated neighborhood size and neural network depth are two completely orthogonal aspects of graph representation, several methods focus on summarizing the neighborhood by aggregating K-hop neighborhoods of nodes while using shallow neural networks. However, these methods still encounter oversmoothing, and suffer from high computation and storage costs. In this paper, we use a modified Markov Diffusion Kernel to derive a variant of GCN called Simple Spectral Graph Convolution (SSGC). Our spectral analysis shows that our simple spectral graph convolution used in SSGC is a trade-off of low- and high-pass filter bands which capture the global and local contexts of each node. We provide two theoretical claims which demonstrate that we can aggregate over a sequence of increasingly larger neighborhoods compared to competitors while limiting severe oversmoothing. Our experimental evaluations show that SSGC with a linear learner is competitive in text and node classification tasks. Moreover, SSGC is comparable to other state-of-the-art methods for node clustering and community prediction tasks.

Regularized Inverse Reinforcement Learning

- Wonseok Jeon, Chen-Yang Su, Paul Barde, Thang Doan, Derek Nowrouzezahrai, Joelle Pineau
- abstract@open-review(Spotlight): Inverse Reinforcement Learning (IRL) aims to facilitate a learner's ability to imitate expert behavior by acquiring reward functions that explain the expert's decisions. Regularized IRLapplies strongly convex regularizers to the learner's policy in order to avoid the expert's behavior being rationalized by arbitrary constant rewards, also known as degenerate solutions. We propose tractable solutions, and practical methods to obtain them, for regularized IRL. Current methods are restricted to the maximum-entropy IRL framework, limiting them to Shannon-entropy regularizers, as well as proposing solutions that are intractable in practice. We present theoretical backing for our proposed IRL method's applicability to both discrete and continuous controls, empirically validating our performance on a variety of tasks.

Anatomy of Catastrophic Forgetting: Hidden Representations and Task Semantics

- Vinay Venkatesh Ramasesh, Ethan Dyer, Maithra Raghu
- abstract@open-review(Poster): Catastrophic forgetting is a recurring challenge to developing versatile deep learning models. Despite its ubiquity, there is limited understanding of its connections to neural network (hidden) representations and task semantics. In this paper, we address this important knowledge gap. Through quantitative analysis of neural representations, we find that deeper layers are disproportionately responsible for forgetting, with sequential training resulting in an erasure of earlier task representational subspaces. Methods to mitigate forgetting stabilize these deeper layers, but show diversity on precise effects, with some increasing feature reuse while others store task representations orthogonally, preventing interference. These insights also enable the development of an analytic argument and empirical picture relating forgetting to task semantic similarity, where we find that maximal forgetting occurs for task sequences with intermediate similarity.

On Fast Adversarial Robustness Adaptation in Model-Agnostic Meta-Learning

- Ren Wang, Kaidi Xu, Sijia Liu, Pin-Yu Chen, Tsui-Wei Weng, Chuang Gan, Meng Wang
- abstract@open-review(Poster): Model-agnostic meta-learning (MAML) has emerged as one of the most successful meta-learning techniques in few-shot learning. It enables us to learn a $\textit{meta-initialization}$ of model parameters (that we call $\textit{meta-model}$) to rapidly adapt to new tasks using a small amount of labeled training data. Despite the generalization power of the meta-model, it remains elusive that how $\textit{adversarial robustness}$ can be maintained by MAML in few-shot learning. In addition to generalization, robustness is also desired for a meta-model to defend adversarial examples (attacks). Toward promoting adversarial robustness in MAML, we first study $\textit{when}$ a robustness-promoting regularization should be incorporated, given the fact that MAML adopts a bi-level (fine-tuning vs. meta-update) learning procedure. We show that robustifying the meta-update stage is sufficient to make robustness adapted to the task-specific fine-tuning stage even if the latter uses a standard training protocol. We also make additional justification on the acquired robustness adaptation by peering into the interpretability of neurons' activation maps. Furthermore, we investigate $\textit{how}$ robust regularization can $\textit{efficiently}$ be designed in MAML. We propose a general but easily-optimized robustness-regularized meta-learning framework, which allows the use of unlabeled data augmentation, fast adversarial attack generation, and computationally-light fine-tuning. In particular, we for the first time show that the auxiliary contrastive learning task can enhance the adversarial robustness of MAML. Finally, extensive experiments are conducted to demonstrate the effectiveness of our proposed methods in robust few-shot learning.

What are the Statistical Limits of Offline RL with Linear Function Approximation?

- Ruosong Wang, Dean Foster, Sham M. Kakade
- abstract@open-review(Spotlight): Offline reinforcement learning seeks to utilize offline (observational) data to guide the learning of (causal) sequential decision making strategies. The hope is that offline reinforcement learning coupled with function approximation methods (to deal with the curse of dimensionality) can provide a means to help alleviate the excessive sample complexity burden in modern sequential decision making problems. However, the extent to which this broader approach can be effective is not well understood, where the literature largely consists of sufficient conditions.

This work focuses on the basic question of what are necessary representational and distributional conditions that permit provable sample-efficient offline reinforcement learning. Perhaps surprisingly, our main result shows that even if: i) we have realizability in that the true value function of \emph{every} policy is linear in a given set of features and 2) our off-policy data has good coverage over all features (under a strong spectral condition), any algorithm still (information-theoretically) requires a number of offline samples that is exponential in the problem horizon to non-trivially estimate the value of \emph{any} given policy. Our results highlight that sample-efficient offline policy evaluation is not possible unless significantly stronger conditions hold; such conditions include either having low distribution shift (where the offline data distribution is close to the distribution of the policy to be evaluated) or significantly stronger representational conditions (beyond realizability).

The geometry of integration in text classification RNNs

- Kyle Aitken, Vinay Venkatesh Ramasesh, Ankush Garg, Yuan Cao, David Sussillo, Niru Maheswaranathan
- abstract@open-review(Poster): Despite the widespread application of recurrent neural networks (RNNs), a unified understanding of how RNNs solve particular tasks remains elusive. In particular, it is unclear what dynamical patterns arise in trained RNNs, and how those pat-terns depend on the training dataset or task. This work addresses these questions in the context of text classification, building on earlier work studying the dynamics of binary sentiment-classification networks (Maheswaranathan et al., 2019). We study text-classification tasks beyond the binary case, exploring the dynamics ofRNNs trained on both natural and synthetic datasets. These dynamics, which we find to be both interpretable and low-dimensional, share a common mechanism across architectures and datasets: specifically, these text-classification networks use low-dimensional attractor manifolds to accumulate evidence for each class as they process the text. The dimensionality and geometry of the attractor manifold are determined by the structure of the training dataset, with the dimensionality reflecting the number of scalar quantities the network remembers in order to classify.In categorical classification, for example, we show that this dimensionality is one less than the number of classes. Correlations in the dataset, such as those induced by ordering, can further reduce the dimensionality of the attractor manifold; we show how to predict this reduction using simple word-count statistics computed on the training dataset. To the degree that integration of evidence towards a decision is a common computational primitive, this work continues to lay the foundation for using dynamical systems techniques to study the inner workings of RNNs.

Behavioral Cloning from Noisy Demonstrations

- Fumihiro Sasaki, Ryota Yamashina
- abstract@open-review(Spotlight): We consider the problem of learning an optimal expert behavior policy given noisy demonstrations that contain observations from both optimal and non-optimal expert behaviors. Popular imitation learning algorithms, such as generative adversarial imitation learning, assume that (clear) demonstrations are given from optimal expert policies but not the non-optimal ones, and thus often fail to imitate the optimal expert behaviors given the noisy demonstrations. Prior works that address the problem require (1) learning policies through environment interactions in the same fashion as reinforcement learning, and (2) annotating each demonstration with confidence scores or rankings. However, such environment interactions and annotations in real-world settings take impractically long training time and a significant human effort. In this paper, we propose an imitation learning algorithm to address the problem without any environment interactions and annotations associated with the non-optimal demonstrations. The proposed algorithm learns ensemble policies with a generalized behavioral cloning (BC) objective function where we exploit another policy already learned by BC. Experimental results show that the proposed algorithm can learn behavior policies that are much closer to the optimal policies than ones learned by BC.

Towards Robust Neural Networks via Close-loop Control

- Zhuotong Chen, Qianxiao Li, Zheng Zhang
- abstract@open-review(Poster): Despite their success in massive engineering applications, deep neural networks are vulnerable to various perturbations due to their black-box nature. Recent study has shown that a deep neural network can misclassify the data even if the input data is perturbed by an imperceptible amount. In this paper, we address the robustness issue of neural networks by a novel close-loop control method from the perspective of dynamic systems. Instead of modifying the parameters in a fixed neural network architecture, a close-loop control process is added to generate control signals adaptively for the perturbed or corrupted data. We connect the robustness of neural networks with optimal control using the geometrical information of underlying data to design the control objective. The detailed analysis shows how the embedding manifolds of state trajectory affect error estimation of the proposed method. Our approach can simultaneously maintain the performance on clean data and improve the robustness against many types of data perturbations. It can also further improve the performance of robustly trained neural networks against different perturbations. To the best of our knowledge, this is the first work that improves the robustness of neural networks with close-loop control.

Projected Latent Markov Chain Monte Carlo: Conditional Sampling of Normalizing Flows

- Chris Cannella, Mohammadreza Soltani, Vahid Tarokh
- abstract@open-review(Poster): We introduce Projected Latent Markov Chain Monte Carlo (PL-MCMC), a technique for sampling from the exact conditional distributions learned by normalizing flows. As a conditional sampling method, PL-MCMC enables Monte Carlo Expectation Maximization (MC-EM) training of normalizing flows from incomplete data. Through experimental tests applying normalizing flows to missing data tasks for a variety of data sets, we demonstrate the efficacy of PL-MCMC for conditional sampling from normalizing flows.

How Does Mixup Help With Robustness and Generalization?

- Linjun Zhang, Zhun Deng, Kenji Kawaguchi, Amirata Ghorbani, James Zou
- abstract@open-review(Spotlight): Mixup is a popular data augmentation technique based on on convex combinations of pairs of examples and their labels. This simple technique has shown to substantially improve both the model's robustness as well as the generalization of the trained model. However, it is not well-understood why such improvement occurs. In this paper, we provide theoretical analysis to demonstrate how using Mixup in training helps model robustness and generalization. For robustness, we show that minimizing the Mixup loss corresponds to approximately minimizing an upper bound of the adversarial loss. This explains why models obtained by Mixup training exhibits robustness to several kinds of adversarial attacks such as Fast Gradient Sign Method (FGSM). For generalization, we prove that Mixup augmentation corresponds to a specific type of data-adaptive regularization which reduces overfitting. Our analysis provides new insights and a framework to understand Mixup.

Understanding the failure modes of out-of-distribution generalization

- Vaishnavh Nagarajan, Anders Andreassen, Behnam Neyshabur
- abstract@open-review(Poster): Empirical studies suggest that machine learning models often rely on features, such as the background, that may be spuriously correlated with the label only during training time, resulting in poor accuracy during test-time. In this work, we identify the fundamental factors that give rise to this behavior, by explaining why models fail this way even in easy-to-learn tasks where one would expect these models to succeed. In particular, through a theoretical study of gradient-descent-trained linear classifiers on some easy-to-learn tasks, we uncover two complementary failure modes. These modes arise from how spurious correlations induce two kinds of skews in the data: one geometric in nature and another, statistical. Finally, we construct natural modifications of image classification datasets to understand when these failure modes can arise in practice. We also design experiments to isolate the two failure modes when training modern neural networks on these datasets.

Usable Information and Evolution of Optimal Representations During Training

- Michael Kleinman, Alessandro Achille, Daksh Idnani, Jonathan Kao
- abstract@open-review(Poster): We introduce a notion of usable information contained in the representation learned by a deep network, and use it to study how optimal representations for the task emerge during training. We show that the implicit regularization coming from training with Stochastic Gradient Descent with a high learning-rate and small batch size plays an important role in learning minimal sufficient representations for the task. In the process of arriving at a minimal sufficient representation, we find that the content of the representation changes dynamically during training. In particular, we find that semantically meaningful but ultimately irrelevant information is encoded in the early transient dynamics of training, before being later discarded. In addition, we evaluate how perturbing the initial part of training impacts the learning dynamics and the resulting representations. We show these effects on both perceptual decision-making tasks inspired by neuroscience literature, as well as on standard image classification tasks.

Adaptive Extra-Gradient Methods for Min-Max Optimization and Games

- Kimon Antonakopoulos, Veronica Belmega, Panayotis Mertikopoulos
- abstract@open-review(Poster): We present a new family of min-max optimization algorithms that automatically exploit the geometry of the gradient data observed at earlier iterations to perform more informative extra-gradient steps in later ones. Thanks to this adaptation mechanism, the proposed method automatically detects whether the problem is smooth or not, without requiring any prior tuning by the optimizer. As a result, the algorithm simultaneously achieves order-optimal convergence rates, \ie it converges to an ε-optimal solution within $\mathcal{O}(1/\varepsilon)$ iterations in smooth problems, and within $\mathcal{O}(1/\varepsilon^2)$ iterations in non-smooth ones. Importantly, these guarantees do not require any of the standard boundedness or Lipschitz continuity conditions that are typically assumed in the literature; in particular, they apply even to problems with singularities (such as resource allocation problems and the like). This adaptation is achieved through the use of a geometric apparatus based on Finsler metrics and a suitably chosen mirror-prox template that allows us to derive sharp convergence rates for the methods at hand.

Emergent Symbols through Binding in External Memory

- Taylor Whittington Webb, Ishan Sinha, Jonathan Cohen

- abstract@open-review(Spotlight): A key aspect of human intelligence is the ability to infer abstract rules directly from high-dimensional sensory data, and to do so given only a limited amount of training experience. Deep neural network algorithms have proven to be a powerful tool for learning directly from high-dimensional data, but currently lack this capacity for data-efficient induction of abstract rules, leading some to argue that symbol-processing mechanisms will be necessary to account for this capacity. In this work, we take a step toward bridging this gap by introducing the Emergent Symbol Binding Network (ESBN), a recurrent network augmented with an external memory that enables a form of variable-binding and indirection. This binding mechanism allows symbol-like representations to emerge through the learning process without the need to explicitly incorporate symbol-processing machinery, enabling the ESBN to learn rules in a manner that is abstracted away from the particular entities to which those rules apply. Across a series of tasks, we show that this architecture displays nearly perfect generalization of learned rules to novel entities given only a limited number of training examples, and outperforms a number of other competitive neural network architectures.

Shapley explainability on the data manifold

- Christopher Frye, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, Ilya Feige
- abstract@open-review(Poster): Explainability in AI is crucial for model development, compliance with regulation, and providing operational nuance to predictions. The Shapley framework for explainability attributes a model's predictions to its input features in a mathematically principled and model-agnostic way. However, general implementations of Shapley explainability make an untenable assumption: that the model's features are uncorrelated. In this work, we demonstrate unambiguous drawbacks of this assumption and develop two solutions to Shapley explainability that respect the data manifold. One solution, based on generative modelling, provides flexible access to data imputations; the other directly learns the Shapley value-function, providing performance and stability at the cost of flexibility. While "off-manifold" Shapley values can (i) give rise to incorrect explanations, (ii) hide implicit model dependence on sensitive attributes, and (iii) lead to unintelligible explanations in higher-dimensional data, on-manifold explainability overcomes these problems.

Reinforcement Learning with Random Delays

- Yann Bouteiller, Simon Ramstedt, Giovanni Beltrame, Christopher Pal, Jonathan Binas
- abstract@open-review(Poster): Action and observation delays commonly occur in many Reinforcement Learning applications, such as remote control scenarios. We study the anatomy of randomly delayed environments, and show that partially resampling trajectory fragments in hindsight allows for off-policy multi-step value estimation. We apply this principle to derive Delay-Correcting Actor-Critic (DCAC), an algorithm based on Soft Actor-Critic with significantly better performance in environments with delays. This is shown theoretically and also demonstrated practically on a delay-augmented version of the MuJoCo continuous control benchmark.

Individually Fair Gradient Boosting

- Alexander Vargo, Fan Zhang, Mikhail Yurochkin, Yuekai Sun
- abstract@open-review(Spotlight): We consider the task of enforcing individual fairness in gradient boosting. Gradient boosting is a popular method for machine learning from tabular data, which arise often in applications where algorithmic fairness is a concern. At a high level, our approach is a functional gradient descent on a (distributionally) robust loss function that encodes our intuition of algorithmic fairness for the ML task at hand. Unlike prior approaches to individual fairness that only work with smooth ML models, our approach also works with non-smooth models such as decision trees. We show that our algorithm converges globally and generalizes. We also demonstrate the efficacy of our algorithm on three ML problems susceptible to algorithmic bias.

Shape or Texture: Understanding Discriminative Features in CNNs

- Md Amirul Islam, Matthew Kowal, Patrick Esser, Sen Jia, Björn Ommer, Konstantinos G. Derpanis, Neil Bruce
- abstract@open-review(Poster): Contrasting the previous evidence that neurons in the later layers of a Convolutional Neural Network (CNN) respond to complex object shapes, recent studies have shown that CNNs actually exhibit a 'texture bias': given an image with both texture and shape cues (e.g., a stylized image), a CNN is biased towards predicting the category corresponding to the texture. However, these previous studies conduct experiments on the final classification output of the network, and fail to robustly evaluate the bias contained (i) in the latent representations, and (ii) on a per-pixel level. In this paper, we design a series of experiments that overcome these issues. We do this with the goal of better understanding what type of shape information contained in the network is discriminative, where shape information is encoded, as well as when the network learns about object shape during training. We show that a network learns the majority of overall shape information at the first few epochs of training and that this information is largely encoded in the last few layers of a CNN. Finally, we show that the encoding of shape does not imply the encoding of localized per-pixel semantic information. The experimental results and findings provide a more accurate understanding of the behaviour of current CNNs, thus helping to inform future design choices.

NOVAS: Non-convex Optimization via Adaptive Stochastic Search for End-to-end Learning and Control

- Ioannis Exarchos, Marcus Aloysius Pereira, Ziyi Wang, Evangelos Theodorou
- abstract@open-review(Poster): In this work we propose the use of adaptive stochastic search as a building block for general, non-convex optimization operations within deep neural network architectures. Specifically, for an objective function located at some layer in the network and parameterized by some network parameters, we employ adaptive stochastic search to perform optimization over its output. This operation is differentiable and does not obstruct the passing of gradients during backpropagation, thus enabling us to incorporate it as a component in end-to-end learning. We study the proposed optimization module's properties and benchmark it against two existing alternatives on a synthetic energy-based structured prediction task, and further showcase its use in stochastic optimal control applications.

SenSeI: Sensitive Set Invariance for Enforcing Individual Fairness

- Mikhail Yurochkin, Yuekai Sun
- abstract@open-review(Oral): In this paper, we cast fair machine learning as invariant machine learning. We first formulate a version of individual fairness that enforces invariance on certain sensitive sets. We then design a transport-based regularizer that enforces this version of individual fairness and develop an algorithm to minimize the regularizer efficiently. Our theoretical results guarantee the proposed approach trains certifiably fair ML models. Finally, in the experimental studies we demonstrate improved fairness metrics in comparison to several recent fair training procedures on three ML tasks that are susceptible to algorithmic bias.

Theoretical Analysis of Self-Training with Deep Networks on Unlabeled Data

- Colin Wei, Kendrick Shen, Yining Chen, Tengyu Ma
- abstract@open-review(Oral): Self-training algorithms, which train a model to fit pseudolabels predicted by another previously-learned model, have been very successful for learning with unlabeled data using neural networks. However, the current theoretical understanding of self-training only applies to linear models. This work provides a unified theoretical analysis of self-training with deep networks for semi-supervised learning, unsupervised domain adaptation, and unsupervised learning. At the core of our analysis is a simple but realistic "expansion" assumption, which states that a low-probability subset of the data must expand to a neighborhood with large probability relative to the subset. We also assume that neighborhoods of examples in different

classes have minimal overlap. We prove that under these assumptions, the minimizers of population objectives based on self-training and input-consistency regularization will achieve high accuracy with respect to ground-truth labels. By using off-the-shelf generalization bounds, we immediately convert this result to sample complexity guarantees for neural nets that are polynomial in the margin and Lipschitzness. Our results help explain the empirical successes of recently proposed self-training algorithms which use input consistency regularization.

Negative Data Augmentation

- Abhishek Sinha, Kumar Ayush, Jiaming Song, Burak Uzkent, Hongxia Jin, Stefano Ermon
- abstract@open-review(Poster): Data augmentation is often used to enlarge datasets with synthetic samples generated in accordance with the underlying data distribution. To enable a wider range of augmentations, we explore negative data augmentation strategies (NDA) that intentionally create out-of-distribution samples. We show that such negative out-of-distribution samples provide information on the support of the data distribution, and can be leveraged for generative modeling and representation learning. We introduce a new GAN training objective where we use NDA as an additional source of synthetic data for the discriminator. We prove that under suitable conditions, optimizing the resulting objective still recovers the true data distribution but can directly bias the generator towards avoiding samples that lack the desired structure. Empirically, models trained with our method achieve improved conditional/unconditional image generation along with improved anomaly detection capabilities. Further, we incorporate the same negative data augmentation strategy in a contrastive learning framework for self-supervised representation learning on images and videos, achieving improved performance on downstream image classification, object detection, and action recognition tasks. These results suggest that prior knowledge on what does not constitute valid data is an effective form of weak supervision across a range of unsupervised learning tasks.

Molecule Optimization by Explainable Evolution

- Binghong Chen, Tianzhe Wang, Chengtao Li, Hanjun Dai, Le Song
- abstract@open-review(Poster): Optimizing molecules for desired properties is a fundamental yet challenging task in chemistry, material science, and drug discovery. This paper develops a novel algorithm for optimizing molecular properties via an Expectation-Maximization (EM) like explainable evolutionary process. The algorithm is designed to mimic human experts in the process of searching for desirable molecules and alternate between two stages: the first stage on explainable local search which identifies rationales, i.e., critical subgraph patterns accounting for desired molecular properties, and the second stage on molecule completion which explores the larger space of molecules containing good rationales. We test our approach against various baselines on a real-world multi-property optimization task where each method is given the same number of queries to the property oracle. We show that our evolution-by-explanation algorithm is 79% better than the best baseline in terms of a generic metric combining aspects such as success rate, novelty, and diversity. Human expert evaluation on optimized molecules shows that 60% of top molecules obtained from our methods are deemed successful.

Estimating Lipschitz constants of monotone deep equilibrium models

- Chirag Pabbaraju, Ezra Winston, J Zico Kolter
- abstract@open-review(Poster): Several methods have been proposed in recent years to provide bounds on the Lipschitz constants of deep networks, which can be used to provide robustness guarantees, generalization bounds, and characterize the smoothness of decision boundaries. However, existing bounds get substantially weaker with increasing depth of the network, which makes it unclear how to apply such bounds to recently proposed models such as the deep equilibrium (DEQ) model, which can be viewed as representing an infinitely-deep network. In this paper, we show that monotone DEQs, a recently-proposed subclass of DEQs, have Lipschitz constants that can be bounded as a simple function of the strong monotonicity parameter of the network. We derive simple-yet-tight bounds on both the input-output mapping and the weight-output mapping defined by these networks, and demonstrate that they are small relative to those for comparable standard DNNs. We show that one can use these bounds to design monotone DEQ models, even with e.g. multi-scale convolutional structure, that still have constraints on the Lipschitz constant. We also highlight how to use these bounds to develop PAC-Bayes generalization bounds that do not depend on any depth of the network, and which avoid the exponential depth-dependence of comparable DNN bounds.

Implicit Gradient Regularization

- David Barrett, Benoit Dherin
- abstract@open-review(Poster): Gradient descent can be surprisingly good at optimizing deep neural networks without overfitting and without explicit regularization. We find that the discrete steps of gradient descent implicitly regularize models by penalizing gradient descent trajectories that have large loss gradients. We call this Implicit Gradient Regularization (IGR) and we use backward error analysis to calculate the size of this regularization. We confirm empirically that implicit gradient regularization biases gradient descent toward flat minima, where test errors are small and solutions are robust to noisy parameter perturbations. Furthermore, we demonstrate that the implicit gradient regularization term can be used as an explicit regularizer, allowing us to control this gradient regularization directly. More broadly, our work indicates that backward error analysis is a useful theoretical approach to the perennial question of how learning rate, model size, and parameter regularization interact to determine the properties of overparameterized models optimized with gradient descent.

Structured Prediction as Translation between Augmented Natural Languages

- Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, RISHITA ANUBHAI, Cicero Nogueira dos Santos, Bing Xiang, Stefano Soatto
- abstract@open-review(Spotlight): We propose a new framework, Translation between Augmented Natural Languages (TANL), to solve many structured prediction language tasks including joint entity and relation extraction, nested named entity recognition, relation classification, semantic role labeling, event extraction, coreference resolution, and dialogue state tracking. Instead of tackling the problem by training task-specific discriminative classifiers, we frame it as a translation task between augmented natural languages, from which the task-relevant information can be easily extracted. Our approach can match or outperform task-specific models on all tasks, and in particular achieves new state-of-the-art results on joint entity and relation extraction (CoNLL04, ADE, NYT, and ACE2005 datasets), relation classification (FewRel and TACRED), and semantic role labeling (CoNLL-2005 and CoNLL-2012). We accomplish this while using the same architecture and hyperparameters for all tasks, and even when training a single model to solve all tasks at the same time (multi-task learning). Finally, we show that our framework can also significantly improve the performance in a low-resource regime, thanks to better use of label semantics.

Faster Binary Embeddings for Preserving Euclidean Distances

- Jinjie Zhang, Rayan Saab
- abstract@open-review(Poster): We propose a fast, distance-preserving, binary embedding algorithm to transform a high-dimensional dataset $\mathcal{T}\subseteq\mathbb{R}^n$ into binary sequences in the cube ${\pm 1}^m$. When $\mathcal{T}$ consists of well-spread (i.e., non-sparse) vectors, our embedding method applies a stable noise-shaping quantization scheme to $A x$ where $A\in\mathbb{R}^{m\times n}$ is a sparse Gaussian random matrix. This contrasts with most binary embedding methods, which usually use $x\mapsto \mathrm{sign}(Ax)$ for the embedding. Moreover, we show that Euclidean distances among the elements of $\mathcal{T}$ are approximated by the ℓ_1 norm on the images of ${\pm 1}^m$ under a fast linear transformation. This again contrasts with standard methods, where the Hamming distance is used instead. Our method is both fast and memory efficient, with time complexity $O(m)$ and space complexity $O(m)$ on well-spread data. When the data is not well-spread, we show that the approach still works provided that data is transformed via a Walsh-Hadamard matrix, but now the cost is $O(n\log n)$ per data point. Further, we prove that the

method is accurate and its associated error is comparable to that of a continuous valued Johnson-Lindenstrauss embedding plus a quantization error that admits a polynomial decay as the embedding dimension m increases. Thus the length of the binary codes required to achieve a desired accuracy is quite small, and we show it can even be compressed further without compromising the accuracy. To illustrate our results, we test the proposed method on natural images and show that it achieves strong performance.

Scalable Transfer Learning with Expert Models

- Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, Cedric Renggli, André Susano Pinto, Sylvain Gelly, Daniel Keysers, Neil Houlsby
- abstract@open-review(Poster): Transfer of pre-trained representations can improve sample efficiency and reduce computational requirements for new tasks. However, representations used for transfer are usually generic, and are not tailored to a particular distribution of downstream tasks. We explore the use of expert representations for transfer with a simple, yet effective, strategy. We train a diverse set of experts by exploiting existing label structures, and use cheap-to-compute performance proxies to select the relevant expert for each target task. This strategy scales the process of transferring to new tasks, since it does not revisit the pre-training data during transfer. Accordingly, it requires little extra compute per target task, and results in a speed-up of 2-3 orders of magnitude compared to competing approaches. Further, we provide an adapter-based architecture able to compress many experts into a single model. We evaluate our approach on two different data sources and demonstrate that it outperforms baselines on over 20 diverse vision tasks in both cases.

Signatory: differentiable computations of the signature and logsignature transforms, on both CPU and GPU

- Patrick Kidger, Terry Lyons
- abstract@open-review(Poster): Signatory is a library for calculating and performing functionality related to the signature and logsignature transforms. The focus is on machine learning, and as such includes features such as CPU parallelism, GPU support, and backpropagation. To our knowledge it is the first GPU-capable library for these operations. Signatory implements new features not available in previous libraries, such as efficient precomputation strategies. Furthermore, several novel algorithmic improvements are introduced, producing substantial real-world speedups even on the CPU without parallelism. The library operates as a Python wrapper around C++, and is compatible with the PyTorch ecosystem. It may be installed directly via \texttt{pip}. Source code, documentation, examples, benchmarks and tests may be found at \texttt{\url{https://github.com/patrick-kidger/signatory}}. The license is Apache-2.0.

Distance-Based Regularisation of Deep Networks for Fine-Tuning

- Henry Gouk, Timothy Hospedales, massimiliano pontil
- abstract@open-review(Poster): We investigate approaches to regularisation during fine-tuning of deep neural networks. First we provide a neural network generalisation bound based on Rademacher complexity that uses the distance the weights have moved from their initial values. This bound has no direct dependence on the number of weights and compares favourably to other bounds when applied to convolutional networks. Our bound is highly relevant for fine-tuning, because providing a network with a good initialisation based on transfer learning means that learning can modify the weights less, and hence achieve tighter generalisation. Inspired by this, we develop a simple yet effective fine-tuning algorithm that constrains the hypothesis class to a small sphere centred on the initial pre-trained weights, thus obtaining provably better generalisation performance than conventional transfer learning. Empirical evaluation shows that our algorithm works well, corroborating our theoretical results. It outperforms both state of the art fine-tuning competitors, and penalty-based alternatives that we show do not directly constrain the radius of the search space.

Decoupling Global and Local Representations via Invertible Generative Flows

- Xuezhe Ma, Xiang Kong, Shanghang Zhang, Eduard H Hovy
- abstract@open-review(Poster): In this work, we propose a new generative model that is capable of automatically decoupling global and local representations of images in an entirely unsupervised setting, by embedding a generative flow in the VAE framework to model the decoder. Specifically, the proposed model utilizes the variational auto-encoding framework to learn a (low-dimensional) vector of latent variables to capture the global information of an image, which is fed as a conditional input to a flow-based invertible decoder with architecture borrowed from style transfer literature. Experimental results on standard image benchmarks demonstrate the effectiveness of our model in terms of density estimation, image generation and unsupervised representation learning. Importantly, this work demonstrates that with only architectural inductive biases, a generative model with a likelihood-based objective is capable of learning decoupled representations, requiring no explicit supervision. The code for our model is available at \url{https://github.com/XuezheMax/wolf}.

Understanding Over-parameterization in Generative Adversarial Networks

- Yogesh Balaji, Mohammadmahdi Sajedi, Neha Mukund Kalibhat, Mucong Ding, Dominik Stöger, Mahdi Soltanolkotabi, Soheil Feizi
- abstract@open-review(Poster): A broad class of unsupervised deep learning methods such as Generative Adversarial Networks (GANs) involve training of overparameterized models where the number of parameters of the model exceeds a certain threshold. Indeed, most successful GANs used in practice are trained using overparameterized generator and discriminator networks, both in terms of depth and width. A large body of work in supervised learning have shown the importance of model overparameterization in the convergence of the gradient descent (GD) to globally optimal solutions. In contrast, the unsupervised setting and GANs in particular involve non-convex concave mini-max optimization problems that are often trained using Gradient Descent/Ascent (GDA). The role and benefits of model overparameterization in the convergence of GDA to a global saddle point in non-convex concave problems is far less understood. In this work, we present a comprehensive analysis of the importance of model overparameterization in GANs both theoretically and empirically. We theoretically show that in an overparameterized GAN model with a 1-layer neural network generator and a linear discriminator, GDA converges to a global saddle point of the underlying non-convex concave min-max problem. To the best of our knowledge, this is the first result for global convergence of GDA in such settings. Our theory is based on a more general result that holds for a broader class of nonlinear generators and discriminators that obey certain assumptions (including deeper generators and random feature discriminators). Our theory utilizes and builds upon a novel connection with the convergence analysis of linear time-varying dynamical systems which may have broader implications for understanding the convergence behavior of GDA for non-convex concave problems involving overparameterized models. We also empirically study the role of model overparameterization in GANs using several large-scale experiments on CIFAR-10 and Celeb-A datasets. Our experiments show that overparameterization improves the quality of generated samples across various model architectures and datasets. Remarkably, we observe that overparameterization leads to faster and more stable convergence behavior of GDA across the board.

Filtered Inner Product Projection for Crosslingual Embedding Alignment

- Vin Sachidananda, Ziyi Yang, Chenguang Zhu
- abstract@open-review(Poster): Due to widespread interest in machine translation and transfer learning, there are numerous algorithms for mapping multiple embeddings to a shared representation space. Recently, these algorithms have been studied in the setting of bilingual lexicon induction where one seeks to align the embeddings of a source and a target language such that translated word pairs lie close to one another in a common representation space. In this paper, we propose a method, Filtered Inner Product Projection (FIPP), for mapping embeddings to a common representation space. As semantic shifts are pervasive across languages and domains, FIPP first identifies the common geometric structure in both embeddings and then, only on the common structure, aligns the Gram matrices of these embeddings. FIPP is applicable even when the source and target embeddings are of differing dimensionalities. Additionally, FIPP provides computational benefits in ease of implementation and is faster to compute than current approaches. Following the baselines in Glavas et al. 2019, we evaluate FIPP both in the context of bilingual lexicon induction and downstream language tasks. We

show that FIPP outperforms existing methods on the XLING BLI dataset for most language pairs while also providing robust performance across downstream tasks.

Deep Partition Aggregation: Provable Defenses against General Poisoning Attacks

- Alexander Levine, Soheil Feizi
- abstract@open-review(Poster): Adversarial poisoning attacks distort training data in order to corrupt the test-time behavior of a classifier. A provable defense provides a certificate for each test sample, which is a lower bound on the magnitude of any adversarial distortion of the training set that can corrupt the test sample's classification. We propose two novel provable defenses against poisoning attacks: (i) Deep Partition Aggregation (DPA), a certified defense against a general poisoning threat model, defined as the insertion or deletion of a bounded number of samples to the training set --- by implication, this threat model also includes arbitrary distortions to a bounded number of images and/or labels; and (ii) Semi-Supervised DPA (SS-DPA), a certified defense against label-flipping poisoning attacks. DPA is an ensemble method where base models are trained on partitions of the training set determined by a hash function. DPA is related to both subset aggregation, a well-studied ensemble method in classical machine learning, as well as to randomized smoothing, a popular provable defense against evasion (inference) attacks. Our defense against label-flipping poison attacks, SS-DPA, uses a semi-supervised learning algorithm as its base classifier model: each base classifier is trained using the entire unlabeled training set in addition to the labels for a partition. SS-DPA significantly outperforms the existing certified defense for label-flipping attacks (Rosenfeld et al., 2020) on both MNIST and CIFAR-10: provably tolerating, for at least half of test images, over 600 label flips (vs. < 200 label flips) on MNIST and over 300 label flips (vs. 175 label flips) on CIFAR-10. Against general poisoning attacks where no prior certified defenses exists, DPA can certify $\geq$ 50% of test images against over 500 poison image insertions on MNIST, and nine insertions on CIFAR-10. These results establish new state-of-the-art provable defenses against general and label-flipping poison attacks. Code is available at https://github.com/alevine0/DPA

Growing Efficient Deep Networks by Structured Continuous Sparsification

- Xin Yuan, Pedro Henrique Pamplona Savarese, Michael Maire
- abstract@open-review(Oral): We develop an approach to growing deep network architectures over the course of training, driven by a principled combination of accuracy and sparsity objectives. Unlike existing pruning or architecture search techniques that operate on full-sized models or supernet architectures, our method can start from a small, simple seed architecture and dynamically grow and prune both layers and filters. By combining a continuous relaxation of discrete network structure optimization with a scheme for sampling sparse subnetworks, we produce compact, pruned networks, while also drastically reducing the computational expense of training. For example, we achieve 49.7% inference FLOPs and 47.4% training FLOPs savings compared to a baseline ResNet-50 on ImageNet, while maintaining 75.2% top-1 validation accuracy --- all without any dedicated fine-tuning stage. Experiments across CIFAR, ImageNet, PASCAL VOC, and Penn Treebank, with convolutional networks for image classification and semantic segmentation, and recurrent networks for language modeling, demonstrate that we both train faster and produce more efficient networks than competing architecture pruning or search methods.

HalentNet: Multimodal Trajectory Forecasting with Hallucinative Intents

- Deyao Zhu, Mohamed Zahran, Li Erran Li, Mohamed Elhoseiny
- abstract@open-review(Poster): Motion forecasting is essential for making intelligent decisions in robotic navigation. As a result, the multi-agent behavioral prediction has become a core component of modern human-robot interaction applications such as autonomous driving. Due to various intentions and interactions among agents, agent trajectories can have multiple possible futures. Hence, the motion forecasting model's ability to cover possible modes becomes essential to enable accurate prediction. Towards this goal, we introduce HalentNet to better model the future motion distribution in addition to a traditional trajectory regression learning objective by incorporating generative augmentation losses. We model intents with unsupervised discrete random variables whose training is guided by a collaboration between two key signals: A discriminative loss that encourages intents' diversity and a hallucinative loss that explores intent transitions (i.e., mixed intents) and encourages their smoothness. This regulates the neural network behavior to be more accurately predictive on uncertain scenarios due to the active yet careful exploration of possible future agent behavior. Our model's learned representation leads to better and more semantically meaningful coverage of the trajectory distribution. Our experiments show that our method can improve over the state-of-the-art trajectory forecasting benchmarks, including vehicles and pedestrians, for about 20% on average FDE and 50% on road boundary violation rate when predicting 6 seconds future. We also conducted human experiments to show that our predicted trajectories received 39.6% more votes than the runner-up approach and 32.2% more votes than our variant without hallucinative mixed intent loss. The code will be released soon.

INT: An Inequality Benchmark for Evaluating Generalization in Theorem Proving

- Yuhuai Wu, Albert Jiang, Jimmy Ba, Roger Baker Grosse
- abstract@open-review(Poster): In learning-assisted theorem proving, one of the most critical challenges is to generalize to theorems unlike those seen at training time. In this paper, we introduce INT, an INequality Theorem proving benchmark designed to test agents' generalization ability. INT is based on a theorem generator, which provides theoretically infinite data and allows us to measure 6 different types of generalization, each reflecting a distinct challenge, characteristic of automated theorem proving. In addition, provides a fast theorem proving environment with sequence-based and graph-based interfaces, conducive to performing learning-based research. We introduce base-lines with architectures including transformers and graph neural networks (GNNs)for INT. Using INT, we find that transformer-based agents achieve stronger test performance for most of the generalization tasks, despite having much larger out-of-distribution generalization gaps than GNNs. We further find that the addition of Monte Carlo Tree Search (MCTS) at test time helps to prove new theorems.

Bayesian Few-Shot Classification with One-vs-Each Pólya-Gamma Augmented Gaussian Processes

- Jake Snell, Richard Zemel
- abstract@open-review(Poster): Few-shot classification (FSC), the task of adapting a classifier to unseen classes given a small labeled dataset, is an important step on the path toward human-like machine learning. Bayesian methods are well-suited to tackling the fundamental issue of overfitting in the few-shot scenario because they allow practitioners to specify prior beliefs and update those beliefs in light of observed data. Contemporary approaches to Bayesian few-shot classification maintain a posterior distribution over model parameters, which is slow and requires storage that scales with model size. Instead, we propose a Gaussian process classifier based on a novel combination of Pólya-Gamma augmentation and the one-vs-each softmax approximation that allows us to efficiently marginalize over functions rather than model parameters. We demonstrate improved accuracy and uncertainty quantification on both standard few-shot classification benchmarks and few-shot domain transfer tasks.

A Hypergradient Approach to Robust Regression without Correspondence

- Yujia Xie, Yixiu Mao, Simiao Zuo, Hongteng Xu, Xiaojing Ye, Tuo Zhao, Hongyuan Zha
- abstract@open-review(Poster): We consider a regression problem, where the correspondence between the input and output data is not available. Such shuffled data are commonly observed in many real world problems. Take flow cytometry as an example: the measuring instruments are unable to preserve the correspondence between the samples and the measurements. Due to the combinatorial nature of the problem, most of the existing methods are only applicable when the sample size is small, and are limited to linear regression models. To overcome such bottlenecks, we propose a new computational framework --- ROBOT --- for the shuffled regression problem, which is applicable to large data and complex models. Specifically, we propose to formulate regression without correspondence as a continuous optimization problem. Then by exploiting the interaction between the regression model and

the data correspondence, we propose to develop a hypergradient approach based on differentiable programming techniques. Such a hypergradient approach essentially views the data correspondence as an operator of the regression model, and therefore it allows us to find a better descent direction for the model parameters by differentiating through the data correspondence. ROBOT is quite general, and can be further extended to an inexact correspondence setting, where the input and output data are not necessarily exactly aligned. Thorough numerical experiments show that ROBOT achieves better performance than existing methods in both linear and nonlinear regression tasks, including real-world applications such as flow cytometry and multi-object tracking.

C-Learning: Horizon-Aware Cumulative Accessibility Estimation

- Panteha Naderian, Gabriel Loaiza-Ganem, Harry J. Braviner, Anthony L. Caterini, Jesse C. Cresswell, Tong Li, Animesh Garg
- abstract@open-review(Poster): Multi-goal reaching is an important problem in reinforcement learning needed to achieve algorithmic generalization. Despite recent advances in this field, current algorithms suffer from three major challenges: high sample complexity, learning only a single way of reaching the goals, and difficulties in solving complex motion planning tasks. In order to address these limitations, we introduce the concept of cumulative accessibility functions, which measure the reachability of a goal from a given state within a specified horizon. We show that these functions obey a recurrence relation, which enables learning from offline interactions. We also prove that optimal cumulative accessibility functions are monotonic in the planning horizon. Additionally, our method can trade off speed and reliability in goal-reaching by suggesting multiple paths to a single goal depending on the provided horizon. We evaluate our approach on a set of multi-goal discrete and continuous control tasks. We show that our method outperforms state-of-the-art goal-reaching algorithms in success rate, sample complexity, and path optimality. Our code is available at https://github.com/layer6ai-labs/CAE, and additional visualizations can be found at https://sites.google.com/view/learning-cae/.

Minimum Width for Universal Approximation

- Sejun Park, Chulhee Yun, Jaeho Lee, Jinwoo Shin
- abstract@open-review(Spotlight): The universal approximation property of width-bounded networks has been studied as a dual of classical universal approximation results on depth-bounded networks. However, the critical width enabling the universal approximation has not been exactly characterized in terms of the input dimension d_x and the output dimension d_y. In this work, we provide the first definitive result in this direction for networks using the ReLU activation functions: The minimum width required for the universal approximation of the L^p functions is exactly $\max{d_x+1,d_y}$. We also prove that the same conclusion does not hold for the uniform approximation with ReLU, but does hold with an additional threshold activation function. Our proof technique can be also used to derive a tighter upper bound on the minimum width required for the universal approximation using networks with general activation functions.

MIROSTAT: A NEURAL TEXT DECODING ALGORITHM THAT DIRECTLY CONTROLS PERPLEXITY

- Sourya Basu, Govardana Sachitanandam Ramachandran, Nitish Shirish Keskar, Lav R. Varshney
- abstract@open-review(Poster): Neural text decoding algorithms strongly influence the quality of texts generated using language models, but popular algorithms like top-k, top-p (nucleus), and temperature-based sampling may yield texts that have objectionable repetition or incoherence. Although these methods generate high-quality text after ad hoc parameter tuning that depends on the language model and the length of generated text, not much is known about the control they provide over the statistics of the output. This is important, however, since recent reports show that humans prefer when perplexity is neither too much nor too little and since we experimentally show that cross-entropy (log of perplexity) has a near-linear relation with repetition. First, we provide a theoretical analysis of perplexity in top-k, top-p, and temperature sampling, under Zipfian statistics. Then, we use this analysis to design a feedback-based adaptive top-k text decoding algorithm called mirostat that generates text (of any length) with a predetermined target value of perplexity without any tuning. Experiments show that for low values of k and p, perplexity drops significantly with generated text length and leads to excessive repetitions (the boredom trap). Contrarily, for large values of k and p, perplexity increases with generated text length and leads to incoherence (confusion trap). Mirostat avoids both traps. Specifically, we show that setting target perplexity value beyond a threshold yields negligible sentence-level repetitions. Experiments with human raters for fluency, coherence, and quality further verify our findings.

Global optimality of softmax policy gradient with single hidden layer neural networks in the mean-field regime

- Andrea Agazzi, Jianfeng Lu
- abstract@open-review(Poster): We study the problem of policy optimization for infinite-horizon discounted Markov Decision Processes with softmax policy and nonlinear function approximation trained with policy gradient algorithms. We concentrate on the training dynamics in the mean-field regime, modeling e.g. the behavior of wide single hidden layer neural networks, when exploration is encouraged through entropy regularization. The dynamics of these models is established as a Wasserstein gradient flow of distributions in parameter space. We further prove global optimality of the fixed points of this dynamics under mild conditions on their initialization.

The Deep Bootstrap Framework: Good Online Learners are Good Offline Generalizers

- Preetum Nakkiran, Behnam Neyshabur, Hanie Sedghi
- abstract@open-review(Poster): We propose a new framework for reasoning about generalization in deep learning. The core idea is to couple the Real World, where optimizers take stochastic gradient steps on the empirical loss, to an Ideal World, where optimizers take steps on the population loss. This leads to an alternate decomposition of test error into: (1) the Ideal World test error plus (2) the gap between the two worlds. If the gap (2) is universally small, this reduces the problem of generalization in offline learning to the problem of optimization in online learning. We then give empirical evidence that this gap between worlds can be small in realistic deep learning settings, in particular supervised image classification. For example, CNNs generalize better than MLPs on image distributions in the Real World, but this is "because" they optimize faster on the population loss in the Ideal World. This suggests our framework is a useful tool for understanding generalization in deep learning, and lays the foundation for future research in this direction.

A Better Alternative to Error Feedback for Communication-Efficient Distributed Learning

- Samuel Horváth, Peter Richtarik
- abstract@open-review(Poster): Modern large-scale machine learning applications require stochastic optimization algorithms to be implemented on distributed computing systems. A key bottleneck of such systems is the communication overhead for exchanging information across the workers, such as stochastic gradients. Among the many techniques proposed to remedy this issue, one of the most successful is the framework of compressed communication with error feedback (EF). EF remains the only known technique that can deal with the error induced by contractive compressors which are not unbiased, such as Top-K or PowerSGD. In this paper, we propose a new and theoretically and practically better alternative to EF for dealing with contractive compressors. In particular, we propose a construction which can transform any contractive compressor into an induced unbiased compressor. Following this transformation, existing methods able to work with unbiased compressors can be applied. We show that our approach leads to vast improvements over EF, including reduced memory requirements, better communication complexity guarantees and fewer assumptions. We further extend our results to federated learning with partial participation following an arbitrary distribution over the nodes and demonstrate the benefits thereof. We perform several numerical experiments which validate our theoretical findings.

DrNAS: Dirichlet Neural Architecture Search

- Xiangning Chen, Ruochen Wang, Minhao Cheng, Xiaocheng Tang, Cho-Jui Hsieh
- abstract@open-review(Poster): This paper proposes a novel differentiable architecture search method by formulating it into a distribution learning problem. We treat the continuously relaxed architecture mixing weight as random variables, modeled by Dirichlet distribution. With recently developed pathwise derivatives, the Dirichlet parameters can be easily optimized with gradient-based optimizer in an end-to-end manner. This formulation improves the generalization ability and induces stochasticity that naturally encourages exploration in the search space. Furthermore, to alleviate the large memory consumption of differentiable NAS, we propose a simple yet effective progressive learning scheme that enables searching directly on large-scale tasks, eliminating the gap between search and evaluation phases. Extensive experiments demonstrate the effectiveness of our method. Specifically, we obtain a test error of 2.46\% for CIFAR-10, 23.7\% for ImageNet under the mobile setting. On NAS-Bench-201, we also achieve state-of-the-art results on all three datasets and provide insights for the effective design of neural architecture search algorithms.

Offline Model-Based Optimization via Normalized Maximum Likelihood Estimation

- Justin Fu, Sergey Levine
- abstract@open-review(Poster): In this work we consider data-driven optimization problems where one must maximize a function given only queries at a fixed set of points. This problem setting emerges in many domains where function evaluation is a complex and expensive process, such as in the design of materials, vehicles, or neural network architectures. Because the available data typically only covers a small manifold of the possible space of inputs, a principal challenge is to be able to construct algorithms that can reason about uncertainty and out-of-distribution values, since a naive optimizer can easily exploit an estimated model to return adversarial inputs. We propose to tackle the MBO problem by leveraging the normalized maximum-likelihood (NML) estimator, which provides a principled approach to handling uncertainty and out-of-distribution inputs. While in the standard formulation NML is intractable, we propose a tractable approximation that allows us to scale our method to high-capacity neural network models. We demonstrate that our method can effectively optimize high-dimensional design problems in a variety of disciplines such as chemistry, biology, and materials engineering.

Viewmaker Networks: Learning Views for Unsupervised Representation Learning

- Alex Tamkin, Mike Wu, Noah Goodman
- abstract@open-review(Poster): Many recent methods for unsupervised representation learning train models to be invariant to different "views," or distorted versions of an input. However, designing these views requires considerable trial and error by human experts, hindering widespread adoption of unsupervised representation learning methods across domains and modalities. To address this, we propose viewmaker networks: generative models that learn to produce useful views from a given input. Viewmakers are stochastic bounded adversaries: they produce views by generating and then adding an ℓ_p-bounded perturbation to the input, and are trained adversarially with respect to the main encoder network. Remarkably, when pretraining on CIFAR-10, our learned views enable comparable transfer accuracy to the well-tuned SimCLR augmentations---despite not including transformations like cropping or color jitter. Furthermore, our learned views significantly outperform baseline augmentations on speech recordings (+9 points on average) and wearable sensor data (+17 points on average). Viewmaker views can also be combined with handcrafted views: they improve robustness to common image corruptions and can increase transfer performance in cases where handcrafted views are less explored. These results suggest that viewmakers may provide a path towards more general representation learning algorithms---reducing the domain expertise and effort needed to pretrain on a much wider set of domains. Code is available at https://github.com/alextamkin/viewmaker.

Robust Pruning at Initialization

- Soufiane Hayou, Jean-Francois Ton, Arnaud Doucet, Yee Whye Teh
- abstract@open-review(Poster): Overparameterized Neural Networks (NN) display state-of-the-art performance. However, there is a growing need for smaller, energy-efficient, neural networks to be able to use machine learning applications on devices with limited computational resources. A popular approach consists of using pruning techniques. While these techniques have traditionally focused on pruning pre-trained NN (LeCun et al.,1990; Hassibi et al., 1993), recent work by Lee et al. (2018) has shown promising results when pruning at initialization. However, for Deep NNs, such procedures remain unsatisfactory as the resulting pruned networks can be difficult to train and, for instance, they do not prevent one layer from being fully pruned. In this paper, we provide a comprehensive theoretical analysis of Magnitude and Gradient based pruning at initialization and training of sparse architectures. This allows us to propose novel principled approaches which we validate experimentally on a variety of NN architectures.

Single-Photon Image Classification

- Thomas Fischbacher, Luciano Sbaiz
- abstract@open-review(Poster): Quantum Computing based Machine Learning mainly focuses on quantum computing hardware that is experimentally challenging to realize due to requiring quantum gates that operate at very low temperature. We demonstrate the existence of a "quantum computing toy model" that illustrates key aspects of quantum information processing while being experimentally accessible with room temperature optics. Pondering the question of the theoretical classification accuracy performance limit for MNIST (respectively "Fashion-MNIST") classifiers, subject to the constraint that a decision has to be made after detection of the very first photon that passed through an image-filter, we show that a machine learning system that is permitted to use quantum interference on the photon's state can substantially outperform any machine learning system that can not. Specifically, we prove that a "classical" MNIST (respectively "Fashion-MNIST") classifier cannot achieve an accuracy of better than 21.28% (respectively 18.28% for "Fashion-MNIST") if it must make a decision after seeing a single photon falling on one of the 28×28 image pixels of a detector array. We further demonstrate that a classifier that is permitted to employ quantum interference by optically transforming the photon state prior to detection can achieve a classification accuracy of at least 41.27% for MNIST (respectively 36.14% for "Fashion-MNIST"). We show in detail how to train the corresponding quantum state transformation with TensorFlow and also explain how this example can serve as a teaching tool for the measurement process in quantum mechanics.

Can a Fruit Fly Learn Word Embeddings?

- Yuchen Liang, Chaitanya Ryali, Benjamin Hoover, Leopold Grinberg, Saket Navlakha, Mohammed J Zaki, Dmitry Krotov
- abstract@open-review(Poster): The mushroom body of the fruit fly brain is one of the best studied systems in neuroscience. At its core it consists of a population of Kenyon cells, which receive inputs from multiple sensory modalities. These cells are inhibited by the anterior paired lateral neuron, thus creating a sparse high dimensional representation of the inputs. In this work we study a mathematical formalization of this network motif and apply it to learning the correlational structure between words and their context in a corpus of unstructured text, a common natural language processing (NLP) task. We show that this network can learn semantic representations of words and can generate both static and context-dependent word embeddings. Unlike conventional methods (e.g., BERT, GloVe) that use dense representations for word embedding, our algorithm encodes semantic meaning of words and their context in the form of sparse binary hash codes. The quality of the learned representations is evaluated on word similarity analysis, word-sense disambiguation, and document classification. It is shown that not only can the fruit fly network motif achieve performance comparable to existing methods in NLP, but, additionally, it uses only a fraction of the computational resources (shorter training time and smaller memory footprint).

VCNet and Functional Targeted Regularization For Learning Causal Effects of Continuous Treatments

- Lizhen Nie, Mao Ye, qiang liu, Dan Nicolae
- abstract@open-review(Oral): Motivated by the rising abundance of observational data with continuous treatments, we investigate the problem of estimating the average dose-response curve (ADRF). Available parametric methods are limited in their model space, and previous attempts in leveraging

neural network to enhance model expressiveness relied on partitioning continuous treatment into blocks and using separate heads for each block; this however produces in practice discontinuous ADRFs. Therefore, the question of how to adapt the structure and training of neural network to estimate ADRFs remains open. This paper makes two important contributions. First, we propose a novel varying coefficient neural network (VCNet) that improves model expressiveness while preserving continuity of the estimated ADRF. Second, to improve finite sample performance, we generalize targeted regularization to obtain a doubly robust estimator of the whole ADRF curve.

Topology-Aware Segmentation Using Discrete Morse Theory

- Xiaoling Hu, Yusu Wang, Li Fuxin, Dimitris Samaras, Chao Chen
- abstract@open-review(Spotlight): In the segmentation of fine-scale structures from natural and biomedical images, per-pixel accuracy is not the only metric of concern. Topological correctness, such as vessel connectivity and membrane closure, is crucial for downstream analysis tasks. In this paper, we propose a new approach to train deep image segmentation networks for better topological accuracy. In particular, leveraging the power of discrete Morse theory (DMT), we identify global structures, including 1D skeletons and 2D patches, which are important for topological accuracy. Trained with a novel loss based on these global structures, the network performance is significantly improved especially near topologically challenging locations (such as weak spots of connections and membranes). On diverse datasets, our method achieves superior performance on both the DICE score and topological metrics.

A Wigner-Eckart Theorem for Group Equivariant Convolution Kernels

- Leon Lang, Maurice Weiler
- abstract@open-review(Poster): Group equivariant convolutional networks (GCNNs) endow classical convolutional networks with additional symmetry priors, which can lead to a considerably improved performance. Recent advances in the theoretical description of GCNNs revealed that such models can generally be understood as performing convolutions with G-steerable kernels, that is, kernels that satisfy an equivariance constraint themselves. While the G-steerability constraint has been derived, it has to date only been solved for specific use cases - a general characterization of G-steerable kernel spaces is still missing. This work provides such a characterization for the practically relevant case of G being any compact group. Our investigation is motivated by a striking analogy between the constraints underlying steerable kernels on the one hand and spherical tensor operators from quantum mechanics on the other hand. By generalizing the famous Wigner-Eckart theorem for spherical tensor operators, we prove that steerable kernel spaces are fully understood and parameterized in terms of 1) generalized reduced matrix elements, 2) Clebsch-Gordan coefficients, and 3) harmonic basis functions on homogeneous spaces.

Non-asymptotic Confidence Intervals of Off-policy Evaluation: Primal and Dual Bounds

- Yihao Feng, Ziyang Tang, na zhang, qiang liu
- abstract@open-review(Poster): Off-policy evaluation (OPE) is the task of estimating the expected reward of a given policy based on offline data previously collected under different policies. Therefore, OPE is a key step in applying reinforcement learning to real-world domains such as medical treatment, where interactive data collection is expensive or even unsafe. As the observed data tends to be noisy and limited, it is essential to provide rigorous uncertainty quantification, not just a point estimation, when applying OPE to make high stakes decisions. This work considers the problem of constructing non-asymptotic confidence intervals in infinite-horizon off-policy evaluation, which remains a challenging open question. We develop a practical algorithm through a primal-dual optimization-based approach, which leverages the kernel Bellman loss (KBL) of Feng et al. 2019 and a new martingale concentration inequality of KBL applicable to time-dependent data with unknown mixing conditions. Our algorithm makes minimum assumptions on the data and the function class of the Q-function, and works for the behavior-agnostic settings where the data is collected under a mix of arbitrary unknown behavior policies. We present empirical results that clearly demonstrate the advantages of our approach over existing methods.

Model Patching: Closing the Subgroup Performance Gap with Data Augmentation

- Karan Goel, Albert Gu, Yixuan Li, Christopher Re
- abstract@open-review(Poster): Classifiers in machine learning are often brittle when deployed. Particularly concerning are models with inconsistent performance on specific subgroups of a class, e.g., exhibiting disparities in skin cancer classification in the presence or absence of a spurious bandage. To mitigate these performance differences, we introduce model patching, a two-stage framework for improving robustness that encourages the model to be invariant to subgroup differences, and focus on class information shared by subgroups. Model patching first models subgroup features within a class and learns semantic transformations between them, and then trains a classifier with data augmentations that deliberately manipulate subgroup features. We instantiate model patching with CAMEL, which (1) uses a CycleGAN to learn the intra-class, inter-subgroup augmentations, and (2) balances subgroup performance using a theoretically-motivated subgroup consistency regularizer, accompanied by a new robust objective. We demonstrate CAMEL's effectiveness on 3 benchmark datasets, with reductions in robust error of up to 33% relative to the best baseline. Lastly, CAMEL successfully patches a model that fails due to spurious features on a real-world skin cancer dataset.

SOLAR: Sparse Orthogonal Learned and Random Embeddings

- Tharun Medini, Beidi Chen, Anshumali Shrivastava
- abstract@open-review(Poster): Dense embedding models are commonly deployed in commercial search engines, wherein all the document vectors are pre-computed, and near-neighbor search (NNS) is performed with the query vector to find relevant documents. However, the bottleneck of indexing a large number of dense vectors and performing an NNS hurts the query time and accuracy of these models. In this paper, we argue that high-dimensional and ultra-sparse embedding is a significantly superior alternative to dense low-dimensional embedding for both query efficiency and accuracy. Extreme sparsity eliminates the need for NNS by replacing them with simple lookups, while its high dimensionality ensures that the embeddings are informative even when sparse. However, learning extremely high dimensional embeddings leads to blow up in the model size. To make the training feasible, we propose a partitioning algorithm that learns such high dimensional embeddings across multiple GPUs without any communication. This is facilitated by our novel asymmetric mixture of Sparse, Orthogonal, Learned and Random (SOLAR) Embeddings. The label vectors are random, sparse, and near-orthogonal by design, while the query vectors are learned and sparse. We theoretically prove that our way of one-sided learning is equivalent to learning both query and label embeddings. With these unique properties, we can successfully train 500K dimensional SOLAR embeddings for the tasks of searching through 1.6M books and multi-label classification on the three largest public datasets. We achieve superior precision and recall compared to the respective state-of-the-art baselines for each task with up to 10 times faster speed.

Neural representation and generation for RNA secondary structures

- Zichao Yan, William L. Hamilton, Mathieu Blanchette
- abstract@open-review(Poster): Our work is concerned with the generation and targeted design of RNA, a type of genetic macromolecule that can adopt complex structures which influence their cellular activities and functions. The design of large scale and complex biological structures spurs dedicated graph-based deep generative modeling techniques, which represents a key but underappreciated aspect of computational drug discovery. In this work, we investigate the principles behind representing and generating different RNA structural modalities, and propose a flexible framework to jointly embed and generate these molecular structures along with their sequence in a meaningful latent space. Equipped with a deep understanding of RNA molecular structures, our most sophisticated encoding and decoding methods operate on the molecular graph as well as the junction tree hierarchy, integrating strong inductive bias about RNA structural regularity and folding mechanism such that high structural validity, stability and diversity of generated RNAs are

achieved. Also, we seek to adequately organize the latent space of RNA molecular embeddings with regard to the interaction with proteins, and targeted optimization is used to navigate in this latent space to search for desired novel RNA molecules.

A Mathematical Exploration of Why Language Models Help Solve Downstream Tasks

- Nikunj Saunshi, Sadhika Malladi, Sanjeev Arora
- abstract@open-review(Poster): Autoregressive language models, pretrained using large text corpora to do well on next word prediction, have been successful at solving many downstream tasks, even with zero-shot usage. However, there is little theoretical understanding of this success. This paper initiates a mathematical study of this phenomenon for the downstream task of text classification by considering the following questions: (1) What is the intuitive connection between the pretraining task of next word prediction and text classification? (2) How can we mathematically formalize this connection and quantify the benefit of language modeling? For (1), we hypothesize, and verify empirically, that classification tasks of interest can be reformulated as sentence completion tasks, thus making language modeling a meaningful pretraining task. With a mathematical formalization of this hypothesis, we make progress towards (2) and show that language models that are ϵ-optimal in cross-entropy (log-perplexity) learn features that can linearly solve such classification tasks with $\mathcal{O}(\sqrt{\epsilon})$ error, thus demonstrating that doing well on language modeling can be beneficial for downstream tasks. We experimentally verify various assumptions and theoretical findings, and also use insights from the analysis to design a new objective function that performs well on some classification tasks.

EigenGame: PCA as a Nash Equilibrium

- Ian Gemp, Brian McWilliams, Claire Vernade, Thore Graepel
- abstract@open-review(Oral): We present a novel view on principal components analysis as a competitive game in which each approximate eigenvector is controlled by a player whose goal is to maximize their own utility function. We analyze the properties of this PCA game and the behavior of its gradient based updates. The resulting algorithm---which combines elements from Oja's rule with a generalized Gram-Schmidt orthogonalization---is naturally decentralized and hence parallelizable through message passing. We demonstrate the scalability of the algorithm with experiments on large image datasets and neural network activations. We discuss how this new view of PCA as a differentiable game can lead to further algorithmic developments and insights.

Noise or Signal: The Role of Image Backgrounds in Object Recognition

- Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, Aleksander Madry
- abstract@open-review(Poster): We assess the tendency of state-of-the-art object recognition models to depend on signals from image backgrounds. We create a toolkit for disentangling foreground and background signal on ImageNet images, and find that (a) models can achieve non-trivial accuracy by relying on the background alone, (b) models often misclassify images even in the presence of correctly classified foregrounds--up to 88% of the time with adversarially chosen backgrounds, and (c) more accurate models tend to depend on backgrounds less. Our analysis of backgrounds brings us closer to understanding which correlations machine learning models use, and how they determine models' out of distribution performance.

Does enhanced shape bias improve neural network robustness to common corruptions?

- Chaithanya Kumar Mummadi, Ranjitha Subramaniam, Robin Hutmacher, Julien Vitay, Volker Fischer, Jan Hendrik Metzen
- abstract@open-review(Poster): Convolutional neural networks (CNNs) learn to extract representations of complex features, such as object shapes and textures to solve image recognition tasks. Recent work indicates that CNNs trained on ImageNet are biased towards features that encode textures and that these alone are sufficient to generalize to unseen test data from the same distribution as the training data but often fail to generalize to out-of-distribution data. It has been shown that augmenting the training data with different image styles decreases this texture bias in favor of increased shape bias while at the same time improving robustness to common corruptions, such as noise and blur. Commonly, this is interpreted as shape bias increasing corruption robustness. However, this relationship is only hypothesized. We perform a systematic study of different ways of composing inputs based on natural images, explicit edge information, and stylization. While stylization is essential for achieving high corruption robustness, we do not find a clear correlation between shape bias and robustness. We conclude that the data augmentation caused by style-variation accounts for the improved corruption robustness and increased shape bias is only a byproduct.

Long Live the Lottery: The Existence of Winning Tickets in Lifelong Learning

- Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, Zhangyang Wang
- abstract@open-review(Poster): The lottery ticket hypothesis states that a highly sparsified sub-network can be trained in isolation, given the appropriate weight initialization. This paper extends that hypothesis from one-shot task learning, and demonstrates for the first time that such extremely compact and independently trainable sub-networks can be also identified in the lifelong learning scenario, which we call lifelong tickets. We show that the resulting lifelong ticket can further be leveraged to improve the performance of learning over continual tasks. However, it is highly non-trivial to conduct network pruning in the lifelong setting. Two critical roadblocks arise: i) As many tasks now arrive sequentially, finding tickets in a greedy weight pruning fashion will inevitably suffer from the intrinsic bias, that the earlier emerging tasks impact more; ii) As lifelong learning is consistently challenged by catastrophic forgetting, the compact network capacity of tickets might amplify the risk of forgetting. In view of those, we introduce two pruning options, e.g., top-down and bottom-up, for finding lifelong tickets. Compared to the top-down pruning that extends vanilla (iterative) pruning over sequential tasks, we show that the bottom-up one, which can dynamically shrink and (re-)expand model capacity, effectively avoids the undesirable excessive pruning in the early stage. We additionally introduce lottery teaching that further overcomes forgetting via knowledge distillation aided by external unlabeled data. Unifying those ingredients, we demonstrate the existence of very competitive lifelong tickets, e.g., achieving 3-8% of the dense model size with even higher accuracy, compared to strong class-incremental learning baselines on CIFAR-10/CIFAR-100/Tiny-ImageNet datasets. Codes available at https://github.com/VITA-Group/Lifelong-Learning-LTH.

Exploring the Uncertainty Properties of Neural Networks' Implicit Priors in the Infinite-Width Limit

- Ben Adlam, Jaehoon Lee, Lechao Xiao, Jeffrey Pennington, Jasper Snoek
- abstract@open-review(Poster): Modern deep learning models have achieved great success in predictive accuracy for many data modalities. However, their application to many real-world tasks is restricted by poor uncertainty estimates, such as overconfidence on out-of-distribution (OOD) data and ungraceful failing under distributional shift. Previous benchmarks have found that ensembles of neural networks (NNs) are typically the best calibrated models on OOD data. Inspired by this, we leverage recent theoretical advances that characterize the function-space prior of an infinitely-wide NN as a Gaussian process, termed the neural network Gaussian process (NNGP). We use the NNGP with a softmax link function to build a probabilistic model for multi-class classification and marginalize over the latent Gaussian outputs to sample from the posterior. This gives us a better understanding of the implicit prior NNs place on function space and allows a direct comparison of the calibration of the NNGP and its finite-width analogue. We also examine the calibration of previous approaches to classification with the NNGP, which treat classification problems as regression to the one-hot labels. In this case the Bayesian posterior is exact, and we compare several heuristics to generate a categorical distribution over classes. We find these methods are well calibrated under distributional shift. Finally, we consider an infinite-width final layer in conjunction with a pre-trained embedding. This replicates the important practical use case of transfer learning and allows scaling to significantly larger datasets. As well as achieving competitive predictive accuracy, this approach is better calibrated than its finite width analogue.

Hierarchical Autoregressive Modeling for Neural Video Compression

- Ruihan Yang, Yibo Yang, Joseph Marino, Stephan Mandt
- abstract@open-review(Poster): Recent work by Marino et al. (2020) showed improved performance in sequential density estimation by combining masked autoregressive flows with hierarchical latent variable models. We draw a connection between such autoregressive generative models and the task of lossy video compression. Specifically, we view recent neural video compression methods (Lu et al., 2019; Yang et al., 2020b; Agustssonet al., 2020) as instances of a generalized stochastic temporal autoregressive transform, and propose avenues for enhancement based on this insight. Comprehensive evaluations on large-scale video data show improved rate-distortion performance over both state-of-the-art neural and conventional video compression methods.

DeLighT: Deep and Light-weight Transformer

- Sachin Mehta, Marjan Ghazvininejad, Srinivasan Iyer, Luke Zettlemoyer, Hannaneh Hajishirzi
- abstract@open-review(Poster): We introduce a deep and light-weight transformer, DeLighT, that delivers similar or better performance than standard transformer-based models with significantly fewer parameters. DeLighT more efficiently allocates parameters both (1) within each Transformer block using the DeLighT transformation, a deep and light-weight transformation and (2) across blocks using block-wise scaling, that allows for shallower and narrower DeLighT blocks near the input and wider and deeper DeLighT blocks near the output. Overall, DeLighT networks are 2.5 to 4 times deeper than standard transformer models and yet have fewer parameters and operations. Experiments on benchmark machine translation and language modeling tasks show that DeLighT matches or improves the performance of baseline Transformers with 2 to 3 times fewer parameters on average.

Learning A Minimax Optimizer: A Pilot Study

- Jiayi Shen, Xiaohan Chen, Howard Heaton, Tianlong Chen, Jialin Liu, Wotao Yin, Zhangyang Wang
- abstract@open-review(Poster): Solving continuous minimax optimization is of extensive practical interest, yet notoriously unstable and difficult. This paper introduces the learning to optimize(L2O) methodology to the minimax problems for the first time and addresses its accompanying unique challenges. We first present Twin-L2O, the first dedicated minimax L2O method consisting of two LSTMs for updating min and max variables separately. The decoupled design is found to facilitate learning, particularly when the min and max variables are highly asymmetric. Empirical experiments on a variety of minimax problems corroborate the effectiveness of Twin-L2O. We then discuss a crucial concern of Twin-L2O, i.e., its inevitably limited generalizability to unseen optimizees. To address this issue, we present two complementary strategies. Our first solution, Enhanced Twin-L2O, is empirically applicable for general minimax problems, by improving L2O training via leveraging curriculum learning. Our second alternative, called Safeguarded Twin-L2O, is a preliminary theoretical exploration stating that under some strong assumptions, it is possible to theoretically establish the convergence of Twin-L2O. We benchmark our algorithms on several testbed problems and compare against state-of-the-art minimax solvers. The code is available at: https://github.com/VITA-Group/L2O-Minimax.

Fuzzy Tiling Activations: A Simple Approach to Learning Sparse Representations Online

- Yangchen Pan, Kirby Banman, Martha White
- abstract@open-review(Poster): Recent work has shown that sparse representations---where only a small percentage of units are active---can significantly reduce interference. Those works, however, relied on relatively complex regularization or meta-learning approaches, that have only been used offline in a pre-training phase. In this work, we pursue a direction that achieves sparsity by design, rather than by learning. Specifically, we design an activation function that produces sparse representations deterministically by construction, and so is more amenable to online training. The idea relies on the simple approach of binning, but overcomes the two key limitations of binning: zero gradients for the flat regions almost everywhere, and lost precision---reduced discrimination---due to coarse aggregation. We introduce a Fuzzy Tiling Activation (FTA) that provides non-negligible gradients and produces overlap between bins that improves discrimination. We first show that FTA is robust under covariate shift in a synthetic online supervised learning problem, where we can vary the level of correlation and drift. Then we move to the deep reinforcement learning setting and investigate both value-based and policy gradient algorithms that use neural networks with FTAs, in classic discrete control and Mujoco continuous control environments. We show that algorithms equipped with FTAs are able to learn a stable policy faster without needing target networks on most domains.

FedBE: Making Bayesian Model Ensemble Applicable to Federated Learning

- Hong-You Chen, Wei-Lun Chao
- abstract@open-review(Poster): Federated learning aims to collaboratively train a strong global model by accessing users' locally trained models but not their own data. A crucial step is therefore to aggregate local models into a global model, which has been shown challenging when users have non-i.i.d. data. In this paper, we propose a novel aggregation algorithm named FedBE, which takes a Bayesian inference perspective by sampling higher-quality global models and combining them via Bayesian model Ensemble, leading to much robust aggregation. We show that an effective model distribution can be constructed by simply fitting a Gaussian or Dirichlet distribution to the local models. Our empirical studies validate FedBE's superior performance, especially when users' data are not i.i.d. and when the neural networks go deeper. Moreover, FedBE is compatible with recent efforts in regularizing users' model training, making it an easily applicable module: you only need to replace the aggregation method but leave other parts of your federated learning algorithm intact.

Randomized Automatic Differentiation

- Deniz Oktay, Nick McGreivy, Joshua Aduol, Alex Beatson, Ryan P Adams
- abstract@open-review(Oral): The successes of deep learning, variational inference, and many other fields have been aided by specialized implementations of reverse-mode automatic differentiation (AD) to compute gradients of mega-dimensional objectives. The AD techniques underlying these tools were designed to compute exact gradients to numerical precision, but modern machine learning models are almost always trained with stochastic gradient descent. Why spend computation and memory on exact (minibatch) gradients only to use them for stochastic optimization? We develop a general framework and approach for randomized automatic differentiation (RAD), which can allow unbiased gradient estimates to be computed with reduced memory in return for variance. We examine limitations of the general approach, and argue that we must leverage problem specific structure to realize benefits. We develop RAD techniques for a variety of simple neural network architectures, and show that for a fixed memory budget, RAD converges in fewer iterations than using a small batch size for feedforward networks, and in a similar number for recurrent networks. We also show that RAD can be applied to scientific computing, and use it to develop a low-memory stochastic gradient method for optimizing the control parameters of a linear reaction-diffusion PDE representing a fission reactor.

Are Neural Nets Modular? Inspecting Functional Modularity Through Differentiable Weight Masks

- Róbert Csordás, Sjoerd van Steenkiste, Jürgen Schmidhuber
- abstract@open-review(Poster): Neural networks (NNs) whose subnetworks implement reusable functions are expected to offer numerous advantages, including compositionality through efficient recombination of functional building blocks, interpretability, preventing catastrophic interference, etc. Understanding if and how NNs are modular could provide insights into how to improve them. Current inspection methods, however, fail to link modules to their functionality. In this paper, we present a novel method based on learning binary weight masks to identify individual weights and subnets

responsible for specific functions. Using this powerful tool, we contribute an extensive study of emerging modularity in NNs that covers several standard architectures and datasets. We demonstrate how common NNs fail to reuse submodules and offer new insights into the related issue of systematic generalization on language tasks.

Calibration tests beyond classification

- David Widmann, Fredrik Lindsten, Dave Zachariah
- abstract@open-review(Poster): Most supervised machine learning tasks are subject to irreducible prediction errors. Probabilistic predictive models address this limitation by providing probability distributions that represent a belief over plausible targets, rather than point estimates. Such models can be a valuable tool in decision-making under uncertainty, provided that the model output is meaningful and interpretable. Calibrated models guarantee that the probabilistic predictions are neither over- nor under-confident. In the machine learning literature, different measures and statistical tests have been proposed and studied for evaluating the calibration of classification models. For regression problems, however, research has been focused on a weaker condition of calibration based on predicted quantiles for real-valued targets. In this paper, we propose the first framework that unifies calibration evaluation and tests for general probabilistic predictive models. It applies to any such model, including classification and regression models of arbitrary dimension. Furthermore, the framework generalizes existing measures and provides a more intuitive reformulation of a recently proposed framework for calibration in multi-class classification. In particular, we reformulate and generalize the kernel calibration error, its estimators, and hypothesis tests using scalar-valued kernels, and evaluate the calibration of real-valued regression problems.

Meta Back-Translation

- Hieu Pham, Xinyi Wang, Yiming Yang, Graham Neubig
- abstract@open-review(Poster): Back-translation is an effective strategy to improve the performance of Neural Machine Translation~(NMT) by generating pseudo-parallel data. However, several recent works have found that better translation quality in the pseudo-parallel data does not necessarily lead to a better final translation model, while lower-quality but diverse data often yields stronger results instead. In this paper we propose a new way to generate pseudo-parallel data for back-translation that directly optimizes the final model performance. Specifically, we propose a meta-learning framework where the back-translation model learns to match the forward-translation model's gradients on the development data with those on the pseudo-parallel data. In our evaluations in both the standard datasets WMT En-De'14 and WMT En-Fr'14, as well as a multilingual translation setting, our method leads to significant improvements over strong baselines.

Attentional Constellation Nets for Few-Shot Learning

- Weijian Xu, yifan xu, Huaijin Wang, Zhuowen Tu
- abstract@open-review(Poster): The success of deep convolutional neural networks builds on top of the learning of effective convolution operations, capturing a hierarchy of structured features via filtering, activation, and pooling. However, the explicit structured features, e.g. object parts, are not expressive in the existing CNN frameworks. In this paper, we tackle the few-shot learning problem and make an effort to enhance structured features by expanding CNNs with a constellation model, which performs cell feature clustering and encoding with a dense part representation; the relationships among the cell features are further modeled by an attention mechanism. With the additional constellation branch to increase the awareness of object parts, our method is able to attain the advantages of the CNNs while making the overall internal representations more robust in the few-shot learning setting. Our approach attains a significant improvement over the existing methods in few-shot learning on the CIFAR-FS, FC100, and mini-ImageNet benchmarks.

Uncertainty Estimation in Autoregressive Structured Prediction

- Andrey Malinin, Mark Gales
- abstract@open-review(Poster): Uncertainty estimation is important for ensuring safety and robustness of AI systems. While most research in the area has focused on un-structured prediction tasks, limited work has investigated general uncertainty estimation approaches for structured prediction. Thus, this work aims to investigate uncertainty estimation for structured prediction tasks within a single unified and interpretable probabilistic ensemble-based framework. We consider: uncertainty estimation for sequence data at the token-level and complete sequence-level; interpretations for, and applications of, various measures of uncertainty; and discuss both the theoretical and practical challenges associated with obtaining them. This work also provides baselines for token-level and sequence-level error detection, and sequence-level out-of-domain input detection on the WMT'14 English-French and WMT'17 English-German translation and LibriSpeech speech recognition datasets.

Measuring Massive Multitask Language Understanding

- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, Jacob Steinhardt
- abstract@open-review(Poster): We propose a new test to measure a text model's multitask accuracy. The test covers 57 tasks including elementary mathematics, US history, computer science, law, and more. To attain high accuracy on this test, models must possess extensive world knowledge and problem solving ability. We find that while most recent models have near random-chance accuracy, the very largest GPT-3 model improves over random chance by almost 20 percentage points on average. However, on every one of the 57 tasks, the best models still need substantial improvements before they can reach expert-level accuracy. Models also have lopsided performance and frequently do not know when they are wrong. Worse, they still have near-random accuracy on some socially important subjects such as morality and law. By comprehensively evaluating the breadth and depth of a model's academic and professional understanding, our test can be used to analyze models across many tasks and to identify important shortcomings.

No MCMC for me: Amortized sampling for fast and stable training of energy-based models

- Will Sussman Grathwohl, Jacob Jin Kelly, Milad Hashemi, Mohammad Norouzi, Kevin Swersky, David Duvenaud
- abstract@open-review(Poster): Energy-Based Models (EBMs) present a flexible and appealing way to represent uncertainty. Despite recent advances, training EBMs on high-dimensional data remains a challenging problem as the state-of-the-art approaches are costly, unstable, and require considerable tuning and domain expertise to apply successfully. In this work, we present a simple method for training EBMs at scale which uses an entropy-regularized generator to amortize the MCMC sampling typically used in EBM training. We improve upon prior MCMC-based entropy regularization methods with a fast variational approximation. We demonstrate the effectiveness of our approach by using it to train tractable likelihood models. Next, we apply our estimator to the recently proposed Joint Energy Model (JEM), where we match the original performance with faster and stable training. This allows us to extend JEM models to semi-supervised classification on tabular data from a variety of continuous domains.

A Distributional Approach to Controlled Text Generation

- Muhammad Khalifa, Hady Elsahar, Marc Dymetman
- abstract@open-review(Oral): We propose a Distributional Approach for addressing Controlled Text Generation from pre-trained Language Models (LM). This approach permits to specify, in a single formal framework, both "pointwise'" and "distributional" constraints over the target LM — to our knowledge, the first model with such generality —while minimizing KL divergence from the initial LM distribution. The optimal target distribution is then uniquely determined as an explicit EBM (Energy-BasedModel) representation. From that optimal representation, we then train a target controlled

Autoregressive LM through an adaptive distributional variant of PolicyGradient. We conduct a first set of experiments over pointwise constraints showing the advantages of our approach over a set of baselines, in terms of obtaining a controlled LM balancing constraint satisfaction with divergence from the pretrained LM. We then perform experiments over distributional constraints, a unique feature of our approach, demonstrating its potential as a remedy to the problem of Bias in Language Models. Through an ablation study, we show the effectiveness of our adaptive technique for obtaining faster convergence. Code available at https://github.com/naver/gdc

Aligning AI With Shared Human Values

- Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, Jacob Steinhardt
- abstract@open-review(Poster): We show how to assess a language model's knowledge of basic concepts of morality. We introduce the ETHICS dataset, a new benchmark that spans concepts in justice, well-being, duties, virtues, and commonsense morality. Models predict widespread moral judgments about diverse text scenarios. This requires connecting physical and social world knowledge to value judgements, a capability that may enable us to steer chatbot outputs or eventually regularize open-ended reinforcement learning agents. With the ETHICS dataset, we find that current language models have a promising but incomplete ability to predict basic human ethical judgements. Our work shows that progress can be made on machine ethics today, and it provides a steppingstone toward AI that is aligned with human values.

Interpreting Graph Neural Networks for NLP With Differentiable Edge Masking

- Michael Sejr Schlichtkrull, Nicola De Cao, Ivan Titov
- abstract@open-review(Spotlight): Graph neural networks (GNNs) have become a popular approach to integrating structural inductive biases into NLP models. However, there has been little work on interpreting them, and specifically on understanding which parts of the graphs (e.g. syntactic trees or co-reference structures) contribute to a prediction. In this work, we introduce a post-hoc method for interpreting the predictions of GNNs which identifies unnecessary edges. Given a trained GNN model, we learn a simple classifier that, for every edge in every layer, predicts if that edge can be dropped. We demonstrate that such a classifier can be trained in a fully differentiable fashion, employing stochastic gates and encouraging sparsity through the expected L_0 norm. We use our technique as an attribution method to analyze GNN models for two tasks -- question answering and semantic role labeling -- providing insights into the information flow in these models. We show that we can drop a large proportion of edges without deteriorating the performance of the model, while we can analyse the remaining edges for interpreting model predictions.

Class Normalization for (Continual)? Generalized Zero-Shot Learning

- Ivan Skorokhodov, Mohamed Elhoseiny
- abstract@open-review(Poster): Normalization techniques have proved to be a crucial ingredient of successful training in a traditional supervised learning regime. However, in the zero-shot learning (ZSL) world, these ideas have received only marginal attention. This work studies normalization in ZSL scenario from both theoretical and practical perspectives. First, we give a theoretical explanation to two popular tricks used in zero-shot learning: normalize+scale and attributes normalization and show that they help training by preserving variance during a forward pass. Next, we demonstrate that they are insufficient to normalize a deep ZSL model and propose Class Normalization (CN): a normalization scheme, which alleviates this issue both provably and in practice. Third, we show that ZSL models typically have more irregular loss surface compared to traditional classifiers and that the proposed method partially remedies this problem. Then, we test our approach on 4 standard ZSL datasets and outperform sophisticated modern SotA with a simple MLP optimized without any bells and whistles and having ~50 times faster training speed. Finally, we generalize ZSL to a broader problem — continual ZSL, and introduce some principled metrics and rigorous baselines for this new setup. The source code is available at https://github.com/universome/class-norm.

Learning explanations that are hard to vary

- Giambattista Parascandolo, Alexander Neitz, ANTONIO ORVIETO, Luigi Gresele, Bernhard Schölkopf
- abstract@open-review(Poster): In this paper, we investigate the principle that good explanations are hard to vary in the context of deep learning. We show that averaging gradients across examples -- akin to a logical OR of patterns -- can favor memorization and `patchwork' solutions that sew together different strategies, instead of identifying invariances. To inspect this, we first formalize a notion of consistency for minima of the loss surface, which measures to what extent a minimum appears only when examples are pooled. We then propose and experimentally validate a simple alternative algorithm based on a logical AND, that focuses on invariances and prevents memorization in a set of real-world tasks. Finally, using a synthetic dataset with a clear distinction between invariant and spurious mechanisms, we dissect learning signals and compare this approach to well-established regularizers.

Degree-Quant: Quantization-Aware Training for Graph Neural Networks

- Shyam Anil Tailor, Javier Fernandez-Marques, Nicholas Donald Lane
- abstract@open-review(Poster): Graph neural networks (GNNs) have demonstrated strong performance on a wide variety of tasks due to their ability to model non-uniform structured data. Despite their promise, there exists little research exploring methods to make them more efficient at inference time. In this work, we explore the viability of training quantized GNNs, enabling the usage of low precision integer arithmetic during inference. For GNNs seemingly unimportant choices in quantization implementation cause dramatic changes in performance. We identify the sources of error that uniquely arise when attempting to quantize GNNs, and propose an architecturally-agnostic and stable method, Degree-Quant, to improve performance over existing quantization-aware training baselines commonly used on other architectures, such as CNNs. We validate our method on six datasets and show, unlike previous quantization attempts, that models generalize to unseen graphs. Models trained with Degree-Quant for INT8 quantization perform as well as FP32 models in most cases; for INT4 models, we obtain up to 26% gains over the baselines. Our work enables up to 4.7x speedups on CPU when using INT8 arithmetic.

Regularization Matters in Policy Optimization - An Empirical Study on Continuous Control

- Zhuang Liu, Xuanlin Li, Bingyi Kang, Trevor Darrell
- abstract@open-review(Spotlight): Deep Reinforcement Learning (Deep RL) has been receiving increasingly more attention thanks to its encouraging performance on a variety of control tasks. Yet, conventional regularization techniques in training neural networks (e.g., L_2 regularization, dropout) have been largely ignored in RL methods, possibly because agents are typically trained and evaluated in the same environment, and because the deep RL community focuses more on high-level algorithm designs. In this work, we present the first comprehensive study of regularization techniques with multiple policy optimization algorithms on continuous control tasks. Interestingly, we find conventional regularization techniques on the policy networks can often bring large improvement, especially on harder tasks. Our findings are shown to be robust against training hyperparameter variations. We also compare these techniques with the more widely used entropy regularization. In addition, we study regularizing different components and find that only regularizing the policy network is typically the best. We further analyze why regularization may help generalization in RL from four perspectives - sample complexity, reward distribution, weight norm, and noise robustness. We hope our study provides guidance for future practices in regularizing policy optimization algorithms. Our code is available at https://github.com/xuanlinli17/iclr2021_rlreg .

Recurrent Independent Mechanisms

- Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, Bernhard Schölkopf
- abstract@open-review(Spotlight): We explore the hypothesis that learning modular structures which reflect the dynamics of the environment can lead to better generalization and robustness to changes that only affect a few of the underlying causes. We propose Recurrent Independent Mechanisms (RIMs), a new recurrent architecture in which multiple groups of recurrent cells operate with nearly independent transition dynamics, communicate only sparingly through the bottleneck of attention, and compete with each other so they are updated only at time steps where they are most relevant. We show that this leads to specialization amongst the RIMs, which in turn allows for remarkably improved generalization on tasks where some factors of variation differ systematically between training and evaluation.

Byzantine-Resilient Non-Convex Stochastic Gradient Descent

- Zeyuan Allen-Zhu, Faeze Ebrahimianghazani, Jerry Li, Dan Alistarh
- abstract@open-review(Poster): We study adversary-resilient stochastic distributed optimization, in which m machines can independently compute stochastic gradients, and cooperate to jointly optimize over their local objective functions. However, an α-fraction of the machines are Byzantine, in that they may behave in arbitrary, adversarial ways. We consider a variant of this procedure in the challenging non-convex case. Our main result is a new algorithm SafeguardSGD, which can provably escape saddle points and find approximate local minima of the non-convex objective. The algorithm is based on a new concentration filtering technique, and its sample and time complexity bounds match the best known theoretical bounds in the stochastic, distributed setting when no Byzantine machines are present.

Our algorithm is very practical: it improves upon the performance of all prior methods when training deep neural networks, it is relatively lightweight, and it is the first method to withstand two recently-proposed Byzantine attacks.

Teaching Temporal Logics to Neural Networks

- Christopher Hahn, Frederik Schmitt, Jens U. Kreber, Markus Norman Rabe, Bernd Finkbeiner
- abstract@open-review(Poster): We study two fundamental questions in neuro-symbolic computing: can deep learning tackle challenging problems in logics end-to-end, and can neural networks learn the semantics of logics. In this work we focus on linear-time temporal logic (LTL), as it is widely used in verification. We train a Transformer on the problem to directly predict a solution, i.e. a trace, to a given LTL formula. The training data is generated with classical solvers, which, however, only provide one of many possible solutions to each formula. We demonstrate that it is sufficient to train on those particular solutions to formulas, and that Transformers can predict solutions even to formulas from benchmarks from the literature on which the classical solver timed out. Transformers also generalize to the semantics of the logics: while they often deviate from the solutions found by the classical solvers, they still predict correct solutions to most formulas.

Bypassing the Ambient Dimension: Private SGD with Gradient Subspace Identification

- Yingxue Zhou, Steven Wu, Arindam Banerjee
- abstract@open-review(Poster): Differentially private SGD (DP-SGD) is one of the most popular methods for solving differentially private empirical risk minimization (ERM). Due to its noisy perturbation on each gradient update, the error rate of DP-SGD scales with the ambient dimension p, the number of parameters in the model. Such dependence can be problematic for over-parameterized models where $p \gg n$, the number of training samples. Existing lower bounds on private ERM show that such dependence on p is inevitable in the worst case. In this paper, we circumvent the dependence on the ambient dimension by leveraging a low-dimensional structure of gradient space in deep networks---that is, the stochastic gradients for deep nets usually stay in a low dimensional subspace in the training process. We propose Projected DP-SGD that performs noise reduction by projecting the noisy gradients to a low-dimensional subspace, which is given by the top gradient eigenspace on a small public dataset. We provide a general sample complexity analysis on the public dataset for the gradient subspace identification problem and demonstrate that under certain low-dimensional assumptions the public sample complexity only grows logarithmically in p. Finally, we provide a theoretical analysis and empirical evaluations to show that our method can substantially improve the accuracy of DP-SGD in the high privacy regime (corresponding to low privacy loss ϵ).

Generative Time-series Modeling with Fourier Flows

- Ahmed Alaa, Alex James Chan, Mihaela van der Schaar
- abstract@open-review(Poster): Generating synthetic time-series data is crucial in various application domains, such as medical prognosis, wherein research is hamstrung by the lack of access to data due to concerns over privacy. Most of the recently proposed methods for generating synthetic time-series rely on implicit likelihood modeling using generative adversarial networks (GANs)—but such models can be difficult to train, and may jeopardize privacy by “memorizing” temporal patterns in training data. In this paper, we propose an explicit likelihood model based on a novel class of normalizing flows that view time-series data in the frequency-domain rather than the time-domain. The proposed flow, dubbed a Fourier flow, uses a discrete Fourier transform (DFT) to convert variable-length time-series with arbitrary sampling periods into fixed-length spectral representations, then applies a (data-dependent) spectral filter to the frequency-transformed time-series. We show that, by virtue of the DFT analytic properties, the Jacobian determinants and inverse mapping for the Fourier flow can be computed efficiently in linearithmic time, without imposing explicit structural constraints as in existing flows such as NICE (Dinh et al. (2014)), RealNVP (Dinh et al. (2016)) and GLOW (Kingma & Dhariwal (2018)). Experiments show that Fourier flows perform competitively compared to state-of-the-art baselines.

Overparameterisation and worst-case generalisation: friend or foe?

- Aditya Krishna Menon, Ankit Singh Rawat, Sanjiv Kumar
- abstract@open-review(Poster): Overparameterised neural networks have demonstrated the remarkable ability to perfectly fit training samples, while still generalising to unseen test samples. However, several recent works have revealed that such models' good average performance does not always translate to good worst-case performance: in particular, they may perform poorly on subgroups that are under-represented in the training set. In this paper, we show that in certain settings, overparameterised models' performance on under-represented subgroups may be improved via post-hoc processing. Specifically, such models' bias can be restricted to their classification layers, and manifest as structured prediction shifts for rare subgroups. We detail two post-hoc correction techniques to mitigate this bias, which operate purely on the outputs of standard model training. We empirically verify that with such post-hoc correction, overparameterisation can improve average and worst-case performance.

Improving Zero-Shot Voice Style Transfer via Disentangled Representation Learning

- Siyang Yuan, Pengyu Cheng, Ruiyi Zhang, Weituo Hao, Zhe Gan, Lawrence Carin
- abstract@open-review(Poster): Voice style transfer, also called voice conversion, seeks to modify one speaker's voice to generate speech as if it came from another (target) speaker. Previous works have made progress on voice conversion with parallel training data and pre-known speakers. However, zero-shot voice style transfer, which learns from non-parallel data and generates voices for previously unseen speakers, remains a challenging problem. In this paper we propose a novel zero-shot voice transfer method via disentangled representation learning. The proposed method first encodes speaker-related style and voice content of each input voice into separate low-dimensional embedding spaces, and then transfers to a new voice by combining the source content embedding and target style embedding through a decoder. With information-theoretic guidance, the style and content embedding spaces are representative and (ideally) independent of each other. On real-world datasets, our method outperforms other baselines and obtains state-of-the-art results in terms of transfer accuracy and voice naturalness.

Meta-learning with negative learning rates

- Alberto Bernacchia
- abstract@open-review(Poster): Deep learning models require a large amount of data to perform well. When data is scarce for a target task, we can transfer the knowledge gained by training on similar tasks to quickly learn the target. A successful approach is meta-learning, or "learning to learn" a distribution of tasks, where "learning" is represented by an outer loop, and "to learn" by an inner loop of gradient descent. However, a number of recent empirical studies argue that the inner loop is unnecessary and more simple models work equally well or even better. We study the performance of MAML as a function of the learning rate of the inner loop, where zero learning rate implies that there is no inner loop. Using random matrix theory and exact solutions of linear models, we calculate an algebraic expression for the test loss of MAML applied to mixed linear regression and nonlinear regression with overparameterized models. Surprisingly, while the optimal learning rate for adaptation is positive, we find that the optimal learning rate for training is always negative, a setting that has never been considered before. Therefore, not only does the performance increase by decreasing the learning rate to zero, as suggested by recent work, but it can be increased even further by decreasing the learning rate to negative values. These results help clarify under what circumstances meta-learning performs best.

Learning to Recombine and Resample Data For Compositional Generalization

- Ekin Akyürek, Afra Feyza Akyürek, Jacob Andreas
- abstract@open-review(Poster): Flexible neural sequence models outperform grammar- and automaton-based counterparts on a variety of tasks. However, neural models perform poorly in settings requiring compositional generalization beyond the training data—particularly to rare or unseen subsequences. Past work has found symbolic scaffolding (e.g. grammars or automata) essential in these settings. We describe R&R, a learned data augmentation scheme that enables a large category of compositional generalizations without appeal to latent symbolic structure. R&R has two components: recombination of original training examples via a prototype-based generative model and resampling of generated examples to encourage extrapolation. Training an ordinary neural sequence model on a dataset augmented with recombined and resampled examples significantly improves generalization in two language processing problems—instruction following (SCAN) and morphological analysis (SIGMORPHON 2018)—where R&R enables learning of new constructions and tenses from as few as eight initial examples.

Dataset Inference: Ownership Resolution in Machine Learning

- Pratyush Maini, Mohammad Yaghini, Nicolas Papernot
- abstract@open-review(Spotlight): With increasingly more data and computation involved in their training, machine learning models constitute valuable intellectual property. This has spurred interest in model stealing, which is made more practical by advances in learning with partial, little, or no supervision. Existing defenses focus on inserting unique watermarks in a model's decision surface, but this is insufficient: the watermarks are not sampled from the training distribution and thus are not always preserved during model stealing. In this paper, we make the key observation that knowledge contained in the stolen model's training set is what is common to all stolen copies. The adversary's goal, irrespective of the attack employed, is always to extract this knowledge or its by-products. This gives the original model's owner a strong advantage over the adversary: model owners have access to the original training data. We thus introduce $\textit{dataset inference}$, the process of identifying whether a suspected model copy has private knowledge from the original model's dataset, as a defense against model stealing. We develop an approach for dataset inference that combines statistical testing with the ability to estimate the distance of multiple data points to the decision boundary. Our experiments on CIFAR10, SVHN, CIFAR100 and ImageNet show that model owners can claim with confidence greater than 99% that their model (or dataset as a matter of fact) was stolen, despite only exposing 50 of the stolen model's training points. Dataset inference defends against state-of-the-art attacks even when the adversary is adaptive. Unlike prior work, it does not require retraining or overfitting the defended model.

Fooling a Complete Neural Network Verifier

- Dániel Zombori, Balázs Bánhelyi, Tibor Csendes, István Megyeri, Márk Jelasity
- abstract@open-review(Poster): The efficient and accurate characterization of the robustness of neural networks to input perturbation is an important open problem. Many approaches exist including heuristic and exact (or complete) methods. Complete methods are expensive but their mathematical formulation guarantees that they provide exact robustness metrics. However, this guarantee is valid only if we assume that the verified network applies arbitrary-precision arithmetic and the verifier is reliable. In practice, however, both the networks and the verifiers apply limited-precision floating point arithmetic. In this paper, we show that numerical roundoff errors can be exploited to craft adversarial networks, in which the actual robustness and the robustness computed by a state-of-the-art complete verifier radically differ. We also show that such adversarial networks can be used to insert a backdoor into any network in such a way that the backdoor is completely missed by the verifier. The attack is easy to detect in its naive form but, as we show, the adversarial network can be transformed to make its detection less trivial. We offer a simple defense against our particular attack based on adding a very small perturbation to the network weights. However, our conjecture is that other numerical attacks are possible, and exact verification has to take into account all the details of the computation executed by the verified networks, which makes the problem significantly harder.

UMEC: Unified model and embedding compression for efficient recommendation systems

- Jiayi Shen, Haotao Wang, Shupeng Gui, Jianchao Tan, Zhangyang Wang, Ji Liu
- abstract@open-review(Poster): The recommendation system (RS) plays an important role in the content recommendation and retrieval scenarios. The core part of the system is the Ranking neural network, which is usually a bottleneck of whole system performance during online inference. In this work, we propose a unified model and embedding compression (UMEC) framework to hammer an efficient neural network-based recommendation system. Our framework jointly learns input feature selection and neural network compression together, and solve them as an end-to-end resource-constrained optimization problem using ADMM. Our method outperforms other baselines in terms of neural network Flops, sparse embedding feature size and the number of sparse embedding features. We evaluate our method on the public benchmark of DLRM, trained over the Kaggle Criteo dataset. The codes can be found at https://github.com/VITA-Group/UMEC.

Evaluations and Methods for Explanation through Robustness Analysis

- Cheng-Yu Hsieh, Chih-Kuan Yeh, Xuanqing Liu, Pradeep Kumar Ravikumar, Seungyeon Kim, Sanjiv Kumar, Cho-Jui Hsieh
- abstract@open-review(Poster): Feature based explanations, that provide importance of each feature towards the model prediction, is arguably one of the most intuitive ways to explain a model. In this paper, we establish a novel set of evaluation criteria for such feature based explanations by robustness analysis. In contrast to existing evaluations which require us to specify some way to "remove" features that could inevitably introduces biases and artifacts, we make use of the subtler notion of smaller adversarial perturbations. By optimizing towards our proposed evaluation criteria, we obtain new explanations that are loosely necessary and sufficient for a prediction. We further extend the explanation to extract the set of features that would move the current prediction to a target class by adopting targeted adversarial attack for the robustness analysis. Through experiments across multiple domains and a user study, we validate the usefulness of our evaluation criteria and our derived explanations.

Learning and Evaluating Representations for Deep One-Class Classification

- Kihyuk Sohn, Chun-Liang Li, Jinsung Yoon, Minho Jin, Tomas Pfister

- abstract@open-review(Poster): We present a two-stage framework for deep one-class classification. We first learn self-supervised representations from one-class data, and then build one-class classifiers on learned representations. The framework not only allows to learn better representations, but also permits building one-class classifiers that are faithful to the target task. We argue that classifiers inspired by the statistical perspective in generative or discriminative models are more effective than existing approaches, such as a normality score from a surrogate classifier. We thoroughly evaluate different self-supervised representation learning algorithms under the proposed framework for one-class classification. Moreover, we present a novel distribution-augmented contrastive learning that extends training distributions via data augmentation to obstruct the uniformity of contrastive representations. In experiments, we demonstrate state-of-the-art performance on visual domain one-class classification benchmarks, including novelty and anomaly detection. Finally, we present visual explanations, confirming that the decision-making process of deep one-class classifiers is intuitive to humans. The code is available at https://github.com/google-research/deep_representation_one_class.

Learning from Protein Structure with Geometric Vector Perceptrons

- Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael John Lamarre Townshend, Ron Dror
- abstract@open-review(Spotlight): Learning on 3D structures of large biomolecules is emerging as a distinct area in machine learning, but there has yet to emerge a unifying network architecture that simultaneously leverages the geometric and relational aspects of the problem domain. To address this gap, we introduce geometric vector perceptrons, which extend standard dense layers to operate on collections of Euclidean vectors. Graph neural networks equipped with such layers are able to perform both geometric and relational reasoning on efficient representations of macromolecules. We demonstrate our approach on two important problems in learning from protein structure: model quality assessment and computational protein design. Our approach improves over existing classes of architectures on both problems, including state-of-the-art convolutional neural networks and graph neural networks. We release our code at https://github.com/drorlab/gvp.

Unsupervised Object Keypoint Learning using Local Spatial Predictability

- Anand Gopalakrishnan, Sjoerd van Steenkiste, Jürgen Schmidhuber
- abstract@open-review(Spotlight): We propose PermaKey, a novel approach to representation learning based on object keypoints. It leverages the predictability of local image regions from spatial neighborhoods to identify salient regions that correspond to object parts, which are then converted to keypoints. Unlike prior approaches, it utilizes predictability to discover object keypoints, an intrinsic property of objects. This ensures that it does not overly bias keypoints to focus on characteristics that are not unique to objects, such as movement, shape, colour etc. We demonstrate the efficacy of PermaKey on Atari where it learns keypoints corresponding to the most salient object parts and is robust to certain visual distractors. Further, on downstream RL tasks in the Atari domain we demonstrate how agents equipped with our keypoints outperform those using competing alternatives, even on challenging environments with moving backgrounds or distractor objects.

Conditional Negative Sampling for Contrastive Learning of Visual Representations

- Mike Wu, Milan Mosse, Chengxu Zhuang, Daniel Yamins, Noah Goodman
- abstract@open-review(Poster): Recent methods for learning unsupervised visual representations, dubbed contrastive learning, optimize the noise-contrastive estimation (NCE) bound on mutual information between two transformations of an image. NCE typically uses randomly sampled negative examples to normalize the objective, but this may often include many uninformative examples either because they are too easy or too hard to discriminate. Taking inspiration from metric learning, we show that choosing semi-hard negatives can yield stronger contrastive representations. To do this, we introduce a family of mutual information estimators that sample negatives conditionally -- in a "ring" around each positive. We prove that these estimators remain lower-bounds of mutual information, with higher bias but lower variance than NCE. Experimentally, we find our approach, applied on top of existing models (IR, CMC, and MoCo) improves accuracy by 2-5% absolute points in each case, measured by linear evaluation on four standard image benchmarks. Moreover, we find continued benefits when transferring features to a variety of new image distributions from the Meta-Dataset collection and to a variety of downstream tasks such as object detection, instance segmentation, and key-point detection.

Communication in Multi-Agent Reinforcement Learning: Intention Sharing

- Woojun Kim, Jongeui Park, Youngchul Sung
- abstract@open-review(Poster): Communication is one of the core components for learning coordinated behavior in multi-agent systems. In this paper, we propose a new communication scheme named Intention Sharing (IS) for multi-agent reinforcement learning in order to enhance the coordination among agents. In the proposed IS scheme, each agent generates an imagined trajectory by modeling the environment dynamics and other agents' actions. The imagined trajectory is the simulated future trajectory of each agent based on the learned model of the environment dynamics and other agents and represents each agent's future action plan. Each agent compresses this imagined trajectory capturing its future action plan to generate its intention message for communication by applying an attention mechanism to learn the relative importance of the components in the imagined trajectory based on the received message from other agents. Numeral results show that the proposed IS scheme outperforms other communication schemes in multi-agent reinforcement learning.

Neural Thompson Sampling

- Weitong ZHANG, Dongruo Zhou, Lihong Li, Quanquan Gu
- abstract@open-review(Poster): Thompson Sampling (TS) is one of the most effective algorithms for solving contextual multi-armed bandit problems. In this paper, we propose a new algorithm, called Neural Thompson Sampling, which adapts deep neural networks for both exploration and exploitation. At the core of our algorithm is a novel posterior distribution of the reward, where its mean is the neural network approximator, and its variance is built upon the neural tangent features of the corresponding neural network. We prove that, provided the underlying reward function is bounded, the proposed algorithm is guaranteed to achieve a cumulative regret of $O(T^{1/2})$, which matches the regret of other contextual bandit algorithms in terms of total round number T. Experimental comparisons with other benchmark bandit algorithms on various data sets corroborate our theory.

Optimal Regularization can Mitigate Double Descent

- Preetum Nakkiran, Prayaag Venkat, Sham M. Kakade, Tengyu Ma
- abstract@open-review(Poster): Recent empirical and theoretical studies have shown that many learning algorithms -- from linear regression to neural networks -- can have test performance that is non-monotonic in quantities such the sample size and model size. This striking phenomenon, often referred to as "double descent", has raised questions of if we need to re-think our current understanding of generalization. In this work, we study whether the double-descent phenomenon can be avoided by using optimal regularization. Theoretically, we prove that for certain linear regression models with isotropic data distribution, optimally-tuned ℓ_2 regularization achieves monotonic test performance as we grow either the sample size or the model size. We also demonstrate empirically that optimally-tuned ℓ_2 regularization can mitigate double descent for more general models, including neural networks. Our results suggest that it may also be informative to study the test risk scalings of various algorithms in the context of appropriately tuned regularization.

Separation and Concentration in Deep Networks

- John Zarka, Florentin Guth, Stéphane Mallat
- abstract@open-review(Poster): Numerical experiments demonstrate that deep neural network classifiers progressively separate class distributions around their mean, achieving linear separability on the training set, and increasing the Fisher discriminant ratio. We explain this mechanism with two types of operators. We prove that a rectifier without biases applied to sign-invariant tight frames can separate class means and increase Fisher ratios. On the opposite, a soft-thresholding on tight frames can reduce within-class variabilities while preserving class means. Variance reduction bounds are proved for Gaussian mixture models. For image classification, we show that separation of class means can be achieved with rectified wavelet tight frames that are not learned. It defines a scattering transform. Learning 1×1 convolutional tight frames along scattering channels and applying a soft-thresholding reduces within-class variabilities. The resulting scattering network reaches the classification accuracy of ResNet-18 on CIFAR-10 and ImageNet, with fewer layers and no learned biases.

Fast Geometric Projections for Local Robustness Certification

- Aymeric Fromherz, Klas Leino, Matt Fredrikson, Bryan Parno, Corina Pasareanu
- abstract@open-review(Spotlight): Local robustness ensures that a model classifies all inputs within an ℓ_p-ball consistently, which precludes various forms of adversarial inputs. In this paper, we present a fast procedure for checking local robustness in feed-forward neural networks with piecewise-linear activation functions. Such networks partition the input space into a set of convex polyhedral regions in which the network's behavior is linear; hence, a systematic search for decision boundaries within the regions around a given input is sufficient for assessing robustness. Crucially, we show how the regions around a point can be analyzed using simple geometric projections, thus admitting an efficient, highly-parallel GPU implementation that excels particularly for the ℓ_2 norm, where previous work has been less effective. Empirically we find this approach to be far more precise than many approximate verification approaches, while at the same time performing multiple orders of magnitude faster than complete verifiers, and scaling to much deeper networks.

Group Equivariant Generative Adversarial Networks

- Neel Dey, Antong Chen, Soheil Ghafurian
- abstract@open-review(Poster): Recent improvements in generative adversarial visual synthesis incorporate real and fake image transformation in a self-supervised setting, leading to increased stability and perceptual fidelity. However, these approaches typically involve image augmentations via additional regularizers in the GAN objective and thus spend valuable network capacity towards approximating transformation equivariance instead of their desired task. In this work, we explicitly incorporate inductive symmetry priors into the network architectures via group-equivariant convolutional networks. Group-convolutions have higher expressive power with fewer samples and lead to better gradient feedback between generator and discriminator. We show that group-equivariance integrates seamlessly with recent techniques for GAN training across regularizers, architectures, and loss functions. We demonstrate the utility of our methods for conditional synthesis by improving generation in the limited data regime across symmetric imaging datasets and even find benefits for natural images with preferred orientation.

InfoBERT: Improving Robustness of Language Models from An Information Theoretic Perspective

- Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, Jingjing Liu
- abstract@open-review(Poster): Large-scale language models such as BERT have achieved state-of-the-art performance across a wide range of NLP tasks. Recent studies, however, show that such BERT-based models are vulnerable facing the threats of textual adversarial attacks. We aim to address this problem from an information-theoretic perspective, and propose InfoBERT, a novel learning framework for robust fine-tuning of pre-trained language models. InfoBERT contains two mutual-information-based regularizers for model training: (i) an Information Bottleneck regularizer, which suppresses noisy mutual information between the input and the feature representation; and (ii) a Robust Feature regularizer, which increases the mutual information between local robust features and global features. We provide a principled way to theoretically analyze and improve the robustness of representation learning for language models in both standard and adversarial training. Extensive experiments demonstrate that InfoBERT achieves state-of-the-art robust accuracy over several adversarial datasets on Natural Language Inference (NLI) and Question Answering (QA) tasks. Our code is available at https://github.com/AI-secure/InfoBERT.

Random Feature Attention

- Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah Smith, Lingpeng Kong
- abstract@open-review(Spotlight): Transformers are state-of-the-art models for a variety of sequence modeling tasks. At their core is an attention function which models pairwise interactions between the inputs at every timestep. While attention is powerful, it does not scale efficiently to long sequences due to its quadratic time and space complexity in the sequence length. We propose RFA, a linear time and space attention that uses random feature methods to approximate the softmax function, and explore its application in transformers. RFA can be used as a drop-in replacement for conventional softmax attention and offers a straightforward way of learning with recency bias through an optional gating mechanism. Experiments on language modeling and machine translation demonstrate that RFA achieves similar or better performance compared to strong transformer baselines. In the machine translation experiment, RFA decodes twice as fast as a vanilla transformer. Compared to existing efficient transformer variants, RFA is competitive in terms of both accuracy and efficiency on three long text classification datasets. Our analysis shows that RFA's efficiency gains are especially notable on long sequences, suggesting that RFA will be particularly useful in tasks that require working with large inputs, fast decoding speed, or low memory footprints.

Using latent space regression to analyze and leverage compositionality in GANs

- Lucy Chai, Jonas Wulff, Phillip Isola
- abstract@open-review(Poster): In recent years, Generative Adversarial Networks have become ubiquitous in both research and public perception, but how GANs convert an unstructured latent code to a high quality output is still an open question. In this work, we investigate regression into the latent space as a probe to understand the compositional properties of GANs. We find that combining the regressor and a pretrained generator provides a strong image prior, allowing us to create composite images from a collage of random image parts at inference time while maintaining global consistency. To compare compositional properties across different generators, we measure the trade-offs between reconstruction of the unrealistic input and image quality of the regenerated samples. We find that the regression approach enables more localized editing of individual image parts compared to direct editing in the latent space, and we conduct experiments to quantify this independence effect. Our method is agnostic to the semantics of edits, and does not require labels or predefined concepts during training. Beyond image composition, our method extends to a number of related applications, such as image inpainting or example-based image editing, which we demonstrate on several GANs and datasets, and because it uses only a single forward pass, it can operate in real-time. Code is available on our project page: https://chail.github.io/latent-composition/.

Go with the flow: Adaptive control for Neural ODEs

- Mathieu Chalvidal, Matthew Ricci, Rufin VanRullen, Thomas Serre
- abstract@open-review(Poster): Despite their elegant formulation and lightweight memory cost, neural ordinary differential equations (NODEs) suffer from known representational limitations. In particular, the single flow learned by NODEs cannot express all homeomorphisms from a given data space to itself, and their static weight parameterization restricts the type of functions they can learn compared to discrete architectures with layer-dependent weights. Here, we describe a new module called neurally-controlled ODE (N-CODE) designed to improve the expressivity of NODEs. The parameters of N-CODE modules are dynamic variables governed by a trainable map from initial or current activation state, resulting in forms of open-loop and closed-

loop control, respectively. A single module is sufficient for learning a distribution on non-autonomous flows that adaptively drive neural representations. We provide theoretical and empirical evidence that N-CODE circumvents limitations of previous NODEs models and show how increased model expressivity manifests in several supervised and unsupervised learning problems. These favorable empirical results indicate the potential of using data- and activity-dependent plasticity in neural networks across numerous domains.

A Learning Theoretic Perspective on Local Explainability

- Jeffrey Li, Vaishnavh Nagarajan, Gregory Plumb, Ameet Talwalkar
- abstract@open-review(Poster): In this paper, we explore connections between interpretable machine learning and learning theory through the lens of local approximation explanations. First, we tackle the traditional problem of performance generalization and bound the test-time predictive accuracy of a model using a notion of how locally explainable it is. Second, we explore the novel problem of explanation generalization which is an important concern for a growing class of finite sample-based local approximation explanations. Finally, we validate our theoretical results empirically and show that they reflect what can be seen in practice.

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby
- abstract@open-review(Oral): While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.

CopulaGNN: Towards Integrating Representational and Correlational Roles of Graphs in Graph Neural Networks

- Jiaqi Ma, Bo Chang, Xuefei Zhang, Qiaozhu Mei
- abstract@open-review(Poster): Graph-structured data are ubiquitous. However, graphs encode diverse types of information and thus play different roles in data representation. In this paper, we distinguish the \textit{representational} and the \textit{correlational} roles played by the graphs in node-level prediction tasks, and we investigate how Graph Neural Network (GNN) models can effectively leverage both types of information. Conceptually, the representational information provides guidance for the model to construct better node features; while the correlational information indicates the correlation between node outcomes conditional on node features. Through a simulation study, we find that many popular GNN models are incapable of effectively utilizing the correlational information. By leveraging the idea of the copula, a principled way to describe the dependence among multivariate random variables, we offer a general solution. The proposed Copula Graph Neural Network (CopulaGNN) can take a wide range of GNN models as base models and utilize both representational and correlational information stored in the graphs. Experimental results on two types of regression tasks verify the effectiveness of the proposed method.

Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability

- Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, Ameet Talwalkar
- abstract@open-review(Poster): We empirically demonstrate that full-batch gradient descent on neural network training objectives typically operates in a regime we call the Edge of Stability. In this regime, the maximum eigenvalue of the training loss Hessian hovers just above the value $2 / \text{(step size)}$, and the training loss behaves non-monotonically over short timescales, yet consistently decreases over long timescales. Since this behavior is inconsistent with several widespread presumptions in the field of optimization, our findings raise questions as to whether these presumptions are relevant to neural network training. We hope that our findings will inspire future efforts aimed at rigorously understanding optimization at the Edge of Stability.

AutoLRS: Automatic Learning-Rate Schedule by Bayesian Optimization on the Fly

- Yuchen Jin, Tianyi Zhou, Liangyu Zhao, Yibo Zhu, Chuanxiong Guo, Marco Canini, Arvind Krishnamurthy
- abstract@open-review(Poster): The learning rate (LR) schedule is one of the most important hyper-parameters needing careful tuning in training DNNs. However, it is also one of the least automated parts of machine learning systems and usually costs significant manual effort and computing. Though there are pre-defined LR schedules and optimizers with adaptive LR, they introduce new hyperparameters that need to be tuned separately for different tasks/datasets. In this paper, we consider the question: Can we automatically tune the LR over the course of training without human involvement? We propose an efficient method, AutoLRS, which automatically optimizes the LR for each training stage by modeling training dynamics. AutoLRS aims to find an LR that minimizes the validation loss, every τ steps. We formulate it as black-box optimization and solve it by Bayesian optimization (BO). However, collecting training instances for BO requires a system to evaluate each LR queried by BO's acquisition function for τ steps, which is prohibitively expensive in practice. Instead, we apply each candidate LR for only $\tau'\ll\tau$ steps and train an exponential model to predict the validation loss after τ steps. This mutual-training process between BO and the exponential model allows us to bound the number of training steps invested in the BO search. We demonstrate the advantages and the generality of AutoLRS through extensive experiments of training DNNs from diverse domains and using different optimizers. The LR schedules auto-generated by AutoLRS leads to a speedup of $1.22\times$, $1.43\times$, and $1.5\times$ when training ResNet-50, Transformer, and BERT, respectively, compared to the LR schedules in their original papers, and an average speedup of $1.31\times$ over state-of-the-art highly tuned LR schedules.

Autoregressive Dynamics Models for Offline Policy Evaluation and Optimization

- Michael R Zhang, Thomas Paine, Ofir Nachum, Cosmin Paduraru, George Tucker, ziyu wang, Mohammad Norouzi
- abstract@open-review(Poster): Standard dynamics models for continuous control make use of feedforward computation to predict the conditional distribution of next state and reward given current state and action using a multivariate Gaussian with a diagonal covariance structure. This modeling choice assumes that different dimensions of the next state and reward are conditionally independent given the current state and action and may be driven by the fact that fully observable physics-based simulation environments entail deterministic transition dynamics. In this paper, we challenge this conditional independence assumption and propose a family of expressive autoregressive dynamics models that generate different dimensions of the next state and reward sequentially conditioned on previous dimensions. We demonstrate that autoregressive dynamics models indeed outperform standard feedforward models in log-likelihood on heldout transitions. Furthermore, we compare different model-based and model-free off-policy evaluation (OPE) methods on RL Unplugged, a suite of offline MuJoCo datasets, and find that autoregressive dynamics models consistently outperform all baselines, achieving a new state-of-the-art. Finally, we show that autoregressive dynamics models are useful for offline policy optimization by serving as a way to enrich the replay buffer through data augmentation and improving performance using model-based planning.

On the role of planning in model-based deep reinforcement learning

- Jessica B Hamrick, Abram L. Friesen, Feryal Behbahani, Arthur Guez, Fabio Viola, Sims Witherspoon, Thomas Anthony, Lars Holger Buesing, Petar Veličković, Theophane Weber
- abstract@open-review(Poster): Model-based planning is often thought to be necessary for deep, careful reasoning and generalization in artificial agents. While recent successes of model-based reinforcement learning (MBRL) with deep function approximation have strengthened this hypothesis, the resulting diversity of model-based methods has also made it difficult to track which components drive success and why. In this paper, we seek to disentangle the contributions of recent methods by focusing on three questions: (1) How does planning benefit MBRL agents? (2) Within planning, what choices drive performance? (3) To what extent does planning improve generalization? To answer these questions, we study the performance of MuZero (Schrittwieser et al., 2019), a state-of-the-art MBRL algorithm with strong connections and overlapping components with many other MBRL algorithms. We perform a number of interventions and ablations of MuZero across a wide range of environments, including control tasks, Atari, and 9x9 Go. Our results suggest the following: (1) Planning is most useful in the learning process, both for policy updates and for providing a more useful data distribution. (2) Using shallow trees with simple Monte-Carlo rollouts is as performant as more complex methods, except in the most difficult reasoning tasks. (3) Planning alone is insufficient to drive strong generalization. These results indicate where and how to utilize planning in reinforcement learning settings, and highlight a number of open questions for future MBRL research.

Zero-Cost Proxies for Lightweight NAS

- Mohamed S Abdelfattah, Abhinav Mehrotra, Łukasz Dudziak, Nicholas Donald Lane
- abstract@open-review(Poster): Neural Architecture Search (NAS) is quickly becoming the standard methodology to design neural network models. However, NAS is typically compute-intensive because multiple models need to be evaluated before choosing the best one. To reduce the computational power and time needed, a proxy task is often used for evaluating each model instead of full training. In this paper, we evaluate conventional reduced-training proxies and quantify how well they preserve ranking between neural network models during search when compared with the rankings produced by final trained accuracy. We propose a series of zero-cost proxies, based on recent pruning literature, that use just a single minibatch of training data to compute a model's score. Our zero-cost proxies use 3 orders of magnitude less computation but can match and even outperform conventional proxies. For example, Spearman's rank correlation coefficient between final validation accuracy and our best zero-cost proxy on NAS-Bench-201 is 0.82, compared to 0.61 for EcoNAS (a recently proposed reduced-training proxy). Finally, we use these zero-cost proxies to enhance existing NAS search algorithms such as random search, reinforcement learning, evolutionary search and predictor-based search. For all search methodologies and across three different NAS datasets, we are able to significantly improve sample efficiency, and thereby decrease computation, by using our zero-cost proxies. For example on NAS-Bench-101, we achieved the same accuracy 4$\times$ quicker than the best previous result. Our code is made public at: https://github.com/mohsaied/zero-cost-nas.

Personalized Federated Learning with First Order Model Optimization

- Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, Jose M. Alvarez
- abstract@open-review(Poster): While federated learning traditionally aims to train a single global model across decentralized local datasets, one model may not always be ideal for all participating clients. Here we propose an alternative, where each client only federates with other relevant clients to obtain a stronger model per client-specific objectives. To achieve this personalization, rather than computing a single model average with constant weights for the entire federation as in traditional FL, we efficiently calculate optimal weighted model combinations for each client, based on figuring out how much a client can benefit from another's model. We do not assume knowledge of any underlying data distributions or client similarities, and allow each client to optimize for arbitrary target distributions of interest, enabling greater flexibility for personalization. We evaluate and characterize our method on a variety of federated settings, datasets, and degrees of local data heterogeneity. Our method outperforms existing alternatives, while also enabling new features for personalized FL such as transfer outside of local data distributions.

Deconstructing the Regularization of BatchNorm

- Yann Dauphin, Ekin Dogus Cubuk
- abstract@open-review(Poster): Batch normalization (BatchNorm) has become a standard technique in deep learning. Its popularity is in no small part due to its often positive effect on generalization. Despite this success, the regularization effect of the technique is still poorly understood. This study aims to decompose BatchNorm into separate mechanisms that are much simpler. We identify three effects of BatchNorm and assess their impact directly with ablations and interventions. Our experiments show that preventing explosive growth at the final layer at initialization and during training can recover a large part of BatchNorm's generalization boost. This regularization mechanism can lift accuracy by 2.9% for Resnet-50 on Imagenet without BatchNorm. We show it is linked to other methods like Dropout and recent initializations like Fixup. Surprisingly, this simple mechanism matches the improvement of 0.9% of the more complex Dropout regularization for the state-of-the-art Efficientnet-B8 model on Imagenet. This demonstrates the underrated effectiveness of simple regularizations and sheds light on directions to further improve generalization for deep nets.

No Cost Likelihood Manipulation at Test Time for Making Better Mistakes in Deep Networks

- Shyamgopal Karthik, Ameya Prabhu, Puneet K. Dokania, Vineet Gandhi
- abstract@open-review(Poster): There has been increasing interest in building deep hierarchy-aware classifiers that aim to quantify and reduce the severity of mistakes, and not just reduce the number of errors. The idea is to exploit the label hierarchy (e.g., the WordNet ontology) and consider graph distances as a proxy for mistake severity. Surprisingly, on examining mistake-severity distributions of the top-1 prediction, we find that current state-of-the-art hierarchy-aware deep classifiers do not always show practical improvement over the standard cross-entropy baseline in making better mistakes. The reason for the reduction in average mistake-severity can be attributed to the increase in low-severity mistakes, which may also explain the noticeable drop in their accuracy. To this end, we use the classical Conditional Risk Minimization (CRM) framework for hierarchy-aware classification. Given a cost matrix and a reliable estimate of likelihoods (obtained from a trained network), CRM simply amends mistakes at inference time; it needs no extra hyperparameters and requires adding just a few lines of code to the standard cross-entropy baseline. It significantly outperforms the state-of-the-art and consistently obtains large reductions in the average hierarchical distance of top-k predictions across datasets, with very little loss in accuracy. CRM, because of its simplicity, can be used with any off-the-shelf trained model that provides reliable likelihood estimates.

Intrinsic-Extrinsic Convolution and Pooling for Learning on 3D Protein Structures

- Pedro Hermosilla, Marco Schäfer, Matej Lang, Gloria Fackelmann, Pere-Pau Vázquez, Barbora Kozlikova, Michael Krone, Tobias Ritschel, Timo Ropinski
- abstract@open-review(Poster): Proteins perform a large variety of functions in living organisms and thus play a key role in biology. However, commonly used algorithms in protein representation learning were not specifically designed for protein data, and are therefore not able to capture all relevant structural levels of a protein during learning. To fill this gap, we propose two new learning operators, specifically designed to process protein structures. First, we introduce a novel convolution operator that considers the primary, secondary, and tertiary structure of a protein by using n-D convolutions defined on both the Euclidean distance, as well as multiple geodesic distances between the atoms in a multi-graph. Second, we introduce a set of hierarchical pooling operators that enable multi-scale protein analysis. We further evaluate the accuracy of our algorithms on common downstream tasks, where we outperform state-of-the-art protein learning algorithms.

Generative Language-Grounded Policy in Vision-and-Language Navigation with Bayes' Rule

- Shuhei Kurita, Kyunghyun Cho
- abstract@open-review(Poster): Vision-and-language navigation (VLN) is a task in which an agent is embodied in a realistic 3D environment and follows an instruction to reach the goal node. While most of the previous studies have built and investigated a discriminative approach, we notice that there are in fact two possible approaches to building such a VLN agent: discriminative and generative. In this paper, we design and investigate a generative language-grounded policy which uses a language model to compute the distribution over all possible instructions i.e. all possible sequences of vocabulary tokens given action and the transition history. In experiments, we show that the proposed generative approach outperforms the discriminative approach in the Room-2-Room (R2R) and Room-4-Room (R4R) datasets, especially in the unseen environments. We further show that the combination of the generative and discriminative policies achieves close to the state-of-the art results in the R2R dataset, demonstrating that the generative and discriminative policies capture the different aspects of VLN.

Learning a Latent Search Space for Routing Problems using Variational Autoencoders

- André Hottung, Bhanu Bhandari, Kevin Tierney
- abstract@open-review(Poster): Methods for automatically learning to solve routing problems are rapidly improving in performance. While most of these methods excel at generating solutions quickly, they are unable to effectively utilize longer run times because they lack a sophisticated search component. We present a learning-based optimization approach that allows a guided search in the distribution of high-quality solutions for a problem instance. More precisely, our method uses a conditional variational autoencoder that learns to map points in a continuous (latent) search space to high-quality, instance-specific routing problem solutions. The learned space can then be searched by any unconstrained continuous optimization method. We show that even using a standard differential evolution search strategy our approach is able to outperform existing purely machine learning based approaches.

Mastering Atari with Discrete World Models

- Danijar Hafner, Timothy P Lillicrap, Mohammad Norouzi, Jimmy Ba
- abstract@open-review(Poster): Intelligent agents need to generalize from past experience to achieve goals in complex environments. World models facilitate such generalization and allow learning behaviors from imagined outcomes to increase sample-efficiency. While learning world models from image inputs has recently become feasible for some tasks, modeling Atari games accurately enough to derive successful behaviors has remained an open challenge for many years. We introduce DreamerV2, a reinforcement learning agent that learns behaviors purely from predictions in the compact latent space of a powerful world model. The world model uses discrete representations and is trained separately from the policy. DreamerV2 constitutes the first agent that achieves human-level performance on the Atari benchmark of 55 tasks by learning behaviors inside a separately trained world model. With the same computational budget and wall-clock time, Dreamer V2 reaches 200M frames and surpasses the final performance of the top single-GPU agents IQN and Rainbow. DreamerV2 is also applicable to tasks with continuous actions, where it learns an accurate world model of a complex humanoid robot and solves stand-up and walking from only pixel inputs.

Spatial Dependency Networks: Neural Layers for Improved Generative Image Modeling

- Đorđe Miladinović, Aleksandar Stanić, Stefan Bauer, Jürgen Schmidhuber, Joachim M. Buhmann
- abstract@open-review(Poster): How to improve generative modeling by better exploiting spatial regularities and coherence in images? We introduce a novel neural network for building image generators (decoders) and apply it to variational autoencoders (VAEs). In our spatial dependency networks (SDNs), feature maps at each level of a deep neural net are computed in a spatially coherent way, using a sequential gating-based mechanism that distributes contextual information across 2-D space. We show that augmenting the decoder of a hierarchical VAE by spatial dependency layers considerably improves density estimation over baseline convolutional architectures and the state-of-the-art among the models within the same class. Furthermore, we demonstrate that SDN can be applied to large images by synthesizing samples of high quality and coherence. In a vanilla VAE setting, we find that a powerful SDN decoder also improves learning disentangled representations, indicating that neural architectures play an important role in this task. Our results suggest favoring spatial dependency over convolutional layers in various VAE settings. The accompanying source code is given at https://github.com/djordjemila/sdn.

Efficient Transformers in Reinforcement Learning using Actor-Learner Distillation

- Emilio Parisotto, Russ Salakhutdinov
- abstract@open-review(Poster): Many real-world applications such as robotics provide hard constraints on power and compute that limit the viable model complexity of Reinforcement Learning (RL) agents. Similarly, in many distributed RL settings, acting is done on un-accelerated hardware such as CPUs, which likewise restricts model size to prevent intractable experiment run times. These "actor-latency" constrained settings present a major obstruction to the scaling up of model complexity that has recently been extremely successful in supervised learning. To be able to utilize large model capacity while still operating within the limits imposed by the system during acting, we develop an "Actor-Learner Distillation" (ALD) procedure that leverages a continual form of distillation that transfers learning progress from a large capacity learner model to a small capacity actor model. As a case study, we develop this procedure in the context of partially-observable environments, where transformer models have had large improvements over LSTMs recently, at the cost of significantly higher computational complexity. With transformer models as the learner and LSTMs as the actor, we demonstrate in several challenging memory environments that using Actor-Learner Distillation largely recovers the clear sample-efficiency gains of the transformer learner model while maintaining the fast inference and reduced total training time of the LSTM actor model.

IOT: Instance-wise Layer Reordering for Transformer Structures

- Jinhua Zhu, Lijun Wu, Yingce Xia, Shufang Xie, Tao Qin, Wengang Zhou, Houqiang Li, Tie-Yan Liu
- abstract@open-review(Poster): With sequentially stacked self-attention, (optional) encoder-decoder attention, and feed-forward layers, Transformer achieves big success in natural language processing (NLP), and many variants have been proposed. Currently, almost all these models assume that the \emph{layer order} is fixed and kept the same across data samples. We observe that different data samples actually favor different orders of the layers. Based on this observation, in this work, we break the assumption of the fixed layer order in Transformer and introduce instance-wise layer reordering into model structure. Our Instance-wise Ordered Transformer (IOT) can model variant functions by reordered layers, which enables each sample to select the better one to improve the model performance under the constraint of almost same number of parameters. To achieve this, we introduce a light predictor with negligible parameter and inference cost to decide the most capable and favorable layer order for any input sequence. Experiments on 3 tasks (neural machine translation, abstractive summarization, and code generation) and 9 datasets demonstrate consistent improvements of our method. We further show that our method can also be applied to other architectures beyond Transformer. Our code is released at Github\footnote{\url{https://github.com/instance-wise-ordered-transformer/IOT}}.

Federated Learning via Posterior Averaging: A New Perspective and Practical Algorithms

- Maruan Al-Shedivat, Jennifer Gillenwater, Eric Xing, Afshin Rostamizadeh
- abstract@open-review(Poster): Federated learning is typically approached as an optimization problem, where the goal is to minimize a global loss function by distributing computation across client devices that possess local data and specify different parts of the global objective. We present an alternative perspective and formulate federated learning as a posterior inference problem, where the goal is to infer a global posterior distribution by having client devices each infer the posterior of their local data. While exact inference is often intractable, this perspective provides a principled way to search for global optima in federated settings. Further, starting with the analysis of federated quadratic objectives, we develop a computation- and

communication-efficient approximate posterior inference algorithm—federated posterior averaging (FedPA). Our algorithm uses MCMC for approximate inference of local posteriors on the clients and efficiently communicates their statistics to the server, where the latter uses them to refine a global estimate of the posterior mode. Finally, we show that FedPA generalizes federated averaging (FedAvg), can similarly benefit from adaptive optimizers, and yields state-of-the-art results on four realistic and challenging benchmarks, converging faster, to better optima.

Getting a CLUE: A Method for Explaining Uncertainty Estimates

- Javier Antoran, Umang Bhatt, Tameem Adel, Adrian Weller, José Miguel Hernández-Lobato
- abstract@open-review(Oral): Both uncertainty estimation and interpretability are important factors for trustworthy machine learning systems. However, there is little work at the intersection of these two areas. We address this gap by proposing a novel method for interpreting uncertainty estimates from differentiable probabilistic models, like Bayesian Neural Networks (BNNs). Our method, Counterfactual Latent Uncertainty Explanations (CLUE), indicates how to change an input, while keeping it on the data manifold, such that a BNN becomes more confident about the input's prediction. We validate CLUE through 1) a novel framework for evaluating counterfactual explanations of uncertainty, 2) a series of ablation experiments, and 3) a user study. Our experiments show that CLUE outperforms baselines and enables practitioners to better understand which input patterns are responsible for predictive uncertainty.

Extracting Strong Policies for Robotics Tasks from Zero-Order Trajectory Optimizers

- Cristina Pinneri, Shambhuraj Sawant, Sebastian Blaes, Georg Martius
- abstract@open-review(Poster): Solving high-dimensional, continuous robotic tasks is a challenging optimization problem. Model-based methods that rely on zero-order optimizers like the cross-entropy method (CEM) have so far shown strong performance and are considered state-of-the-art in the model-based reinforcement learning community. However, this success comes at the cost of high computational complexity, being therefore not suitable for real-time control. In this paper, we propose a technique to jointly optimize the trajectory and distill a policy, which is essential for fast execution in real robotic systems. Our method builds upon standard approaches, like guidance cost and dataset aggregation, and introduces a novel adaptive factor which prevents the optimizer from collapsing to the learner's behavior at the beginning of the training. The extracted policies reach unprecedented performance on challenging tasks as making a humanoid stand up and opening a door without reward shaping

Sharpness-aware Minimization for Efficiently Improving Generalization

- Pierre Foret, Ariel Kleiner, Hossein Mobahi, Behnam Neyshabur
- abstract@open-review(Spotlight): In today's heavily overparameterized models, the value of the training loss provides few guarantees on model generalization ability. Indeed, optimizing only the training loss value, as is commonly done, can easily lead to suboptimal model quality. Motivated by the connection between geometry of the loss landscape and generalization---including a generalization bound that we prove here---we introduce a novel, effective procedure for instead simultaneously minimizing loss value and loss sharpness. In particular, our procedure, Sharpness-Aware Minimization (SAM), seeks parameters that lie in neighborhoods having uniformly low loss; this formulation results in a min-max optimization problem on which gradient descent can be performed efficiently. We present empirical results showing that SAM improves model generalization across a variety of benchmark datasets (e.g., CIFAR-{10, 100}, ImageNet, finetuning tasks) and models, yielding novel state-of-the-art performance for several. Additionally, we find that SAM natively provides robustness to label noise on par with that provided by state-of-the-art procedures that specifically target learning with noisy labels.

BREEDS: Benchmarks for Subpopulation Shift

- Shibani Santurkar, Dimitris Tsipras, Aleksander Madry
- abstract@open-review(Poster): We develop a methodology for assessing the robustness of models to subpopulation shift---specifically, their ability to generalize to novel data subpopulations that were not observed during training. Our approach leverages the class structure underlying existing datasets to control the data subpopulations that comprise the training and test distributions. This enables us to synthesize realistic distribution shifts whose sources can be precisely controlled and characterized, within existing large-scale datasets. Applying this methodology to the ImageNet dataset, we create a suite of subpopulation shift benchmarks of varying granularity. We then validate that the corresponding shifts are tractable by obtaining human baselines. Finally, we utilize these benchmarks to measure the sensitivity of standard model architectures as well as the effectiveness of existing train-time robustness interventions.

On the Impossibility of Global Convergence in Multi-Loss Optimization

- Alistair Letcher
- abstract@open-review(Poster): Under mild regularity conditions, gradient-based methods converge globally to a critical point in the single-loss setting. This is known to break down for vanilla gradient descent when moving to multi-loss optimization, but can we hope to build some algorithm with global guarantees? We negatively resolve this open problem by proving that desirable convergence properties cannot simultaneously hold for any algorithm. Our result has more to do with the existence of games with no satisfactory outcomes, than with algorithms per se. More explicitly we construct a two-player game with zero-sum interactions whose losses are both coercive and analytic, but whose only simultaneous critical point is a strict maximum. Any 'reasonable' algorithm, defined to avoid strict maxima, will therefore fail to converge. This is fundamentally different from single losses, where coercivity implies existence of a global minimum. Moreover, we prove that a wide range of existing gradient-based methods almost surely have bounded but non-convergent iterates in a constructed zero-sum game for suitably small learning rates. It nonetheless remains an open question whether such behavior can arise in high-dimensional games of interest to ML practitioners, such as GANs or multi-agent RL.

Efficient Wasserstein Natural Gradients for Reinforcement Learning

- Ted Moskovitz, Michael Arbel, Ferenc Huszar, Arthur Gretton
- abstract@open-review(Poster): A novel optimization approach is proposed for application to policy gradient methods and evolution strategies for reinforcement learning (RL). The procedure uses a computationally efficient \emph{Wasserstein natural gradient} (WNG) descent that takes advantage of the geometry induced by a Wasserstein penalty to speed optimization. This method follows the recent theme in RL of including divergence penalties in the objective to establish trust regions. Experiments on challenging tasks demonstrate improvements in both computational cost and performance over advanced baselines.

Optimal Rates for Averaged Stochastic Gradient Descent under Neural Tangent Kernel Regime

- Atsushi Nitanda, Taiji Suzuki
- abstract@open-review(Oral): We analyze the convergence of the averaged stochastic gradient descent for overparameterized two-layer neural networks for regression problems. It was recently found that a neural tangent kernel (NTK) plays an important role in showing the global convergence of gradient-based methods under the NTK regime, where the learning dynamics for overparameterized neural networks can be almost characterized by that for the associated reproducing kernel Hilbert space (RKHS). However, there is still room for a convergence rate analysis in the NTK regime. In this study, we show that the averaged stochastic gradient descent can achieve the minimax optimal convergence rate, with the global convergence guarantee, by

exploiting the complexities of the target function and the RKHS associated with the NTK. Moreover, we show that the target function specified by the NTK of a ReLU network can be learned at the optimal convergence rate through a smooth approximation of a ReLU network under certain conditions.

PMI-Masking: Principled masking of correlated spans

- Yoav Levine, Barak Lenz, Opher Lieber, Omri Abend, Kevin Leyton-Brown, Moshe Tennenholtz, Yoav Shoham
- abstract@open-review(Spotlight): Masking tokens uniformly at random constitutes a common flaw in the pretraining of Masked Language Models (MLMs) such as BERT. We show that such uniform masking allows an MLM to minimize its training objective by latching onto shallow local signals, leading to pretraining inefficiency and suboptimal downstream performance. To address this flaw, we propose PMI-Masking, a principled masking strategy based on the concept of Pointwise Mutual Information (PMI), which jointly masks a token n-gram if it exhibits high collocation over the corpus. PMI-Masking motivates, unifies, and improves upon prior more heuristic approaches that attempt to address the drawback of random uniform token masking, such as whole-word masking, entity/phrase masking, and random-span masking. Specifically, we show experimentally that PMI-Masking reaches the performance of prior masking approaches in half the training time, and consistently improves performance at the end of pretraining.

Multi-Level Local SGD: Distributed SGD for Heterogeneous Hierarchical Networks

- Timothy Castiglia, Anirban Das, Stacy Patterson
- abstract@open-review(Poster): We propose Multi-Level Local SGD, a distributed stochastic gradient method for learning a smooth, non-convex objective in a multi-level communication network with heterogeneous workers. Our network model consists of a set of disjoint sub-networks, with a single hub and multiple workers; further, workers may have different operating rates. The hubs exchange information with one another via a connected, but not necessarily complete communication network. In our algorithm, sub-networks execute a distributed SGD algorithm, using a hub-and-spoke paradigm, and the hubs periodically average their models with neighboring hubs. We first provide a unified mathematical framework that describes the Multi-Level Local SGD algorithm. We then present a theoretical analysis of the algorithm; our analysis shows the dependence of the convergence error on the worker node heterogeneity, hub network topology, and the number of local, sub-network, and global iterations. We illustrate the effectiveness of our algorithm in a multi-level network with slow workers via simulation-based experiments.

Kanerva++: Extending the Kanerva Machine With Differentiable, Locally Block Allocated Latent Memory

- Jason Ramapuram, Yan Wu, Alexandros Kalousis
- abstract@open-review(Poster): Episodic and semantic memory are critical components of the human memory model. The theory of complementary learning systems (McClelland et al., 1995) suggests that the compressed representation produced by a serial event (episodic memory) is later restructured to build a more generalized form of reusable knowledge (semantic memory). In this work, we develop a new principled Bayesian memory allocation scheme that bridges the gap between episodic and semantic memory via a hierarchical latent variable model. We take inspiration from traditional heap allocation and extend the idea of locally contiguous memory to the Kanerva Machine, enabling a novel differentiable block allocated latent memory. In contrast to the Kanerva Machine, we simplify the process of memory writing by treating it as a fully feed forward deterministic process, relying on the stochasticity of the read key distribution to disperse information within the memory. We demonstrate that this allocation scheme improves performance in memory conditional image generation, resulting in new state-of-the-art conditional likelihood values on binarized MNIST (≤41.58 nats/image) , binarized Omniglot (≤66.24 nats/image), as well as presenting competitive performance on CIFAR10, DMLab Mazes, Celeb-A and ImageNet32×32.

EVALUATION OF NEURAL ARCHITECTURES TRAINED WITH SQUARE LOSS VS CROSS-ENTROPY IN CLASSIFICATION TASKS

- Like Hui, Mikhail Belkin
- abstract@open-review(Poster): Modern neural architectures for classification tasks are trained using the cross-entropy loss, which is widely believed to be empirically superior to the square loss. In this work we provide evidence indicating that this belief may not be well-founded. We explore several major neural architectures and a range of standard benchmark datasets for NLP, automatic speech recognition (ASR) and computer vision tasks to show that these architectures, with the same hyper-parameter settings as reported in the literature, perform comparably or better when trained with the square loss, even after equalizing computational resources. Indeed, we observe that the square loss produces better results in the dominant majority of NLP and ASR experiments. Cross-entropy appears to have a slight edge on computer vision tasks.

We argue that there is little compelling empirical or theoretical evidence indicating a clear-cut advantage to the cross-entropy loss. Indeed, in our experiments, performance on nearly all non-vision tasks can be improved, sometimes significantly, by switching to the square loss. Furthermore, training with square loss appears to be less sensitive to the randomness in initialization. We posit that training using the square loss for classification needs to be a part of best practices of modern deep learning on equal footing with cross-entropy.

Self-supervised Learning from a Multi-view Perspective

- Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, Louis-Philippe Morency
- abstract@open-review(Poster): As a subset of unsupervised representation learning, self-supervised representation learning adopts self-defined signals as supervision and uses the learned representation for downstream tasks, such as object detection and image captioning. Many proposed approaches for self-supervised learning follow naturally a multi-view perspective, where the input (e.g., original images) and the self-supervised signals (e.g., augmented images) can be seen as two redundant views of the data. Building from this multi-view perspective, this paper provides an information-theoretical framework to better understand the properties that encourage successful self-supervised learning. Specifically, we demonstrate that self-supervised learned representations can extract task-relevant information and discard task-irrelevant information. Our theoretical framework paves the way to a larger space of self-supervised learning objective design. In particular, we propose a composite objective that bridges the gap between prior contrastive and predictive learning objectives, and introduce an additional objective term to discard task-irrelevant information. To verify our analysis, we conduct controlled experiments to evaluate the impact of the composite objectives. We also explore our framework's empirical generalization beyond the multi-view perspective, where the cross-view redundancy may not be clearly observed.

Interpretable Neural Architecture Search via Bayesian Optimisation with Weisfeiler-Lehman Kernels

- Binxin Ru, Xingchen Wan, Xiaowen Dong, Michael Osborne
- abstract@open-review(Poster): Current neural architecture search (NAS) strategies focus only on finding a single, good, architecture. They offer little insight into why a specific network is performing well, or how we should modify the architecture if we want further improvements. We propose a Bayesian optimisation (BO) approach for NAS that combines the Weisfeiler-Lehman graph kernel with a Gaussian process surrogate. Our method not only optimises the architecture in a highly data-efficient manner, but also affords interpretability by discovering useful network features and their corresponding impact on the network performance. Moreover, our method is capable of capturing the topological structures of the architectures and is scalable to large graphs, thus making the high-dimensional and graph-like search spaces amenable to BO. We demonstrate empirically that our surrogate model is capable of identifying useful motifs which can guide the generation of new architectures. We finally show that our method outperforms existing NAS approaches to achieve the state of the art on both closed- and open-domain search spaces.

CT-Net: Channel Tensorization Network for Video Classification

- Kunchang Li, Xianhang Li, Yali Wang, Jun Wang, Yu Qiao
- abstract@open-review(Poster): 3D convolution is powerful for video classification but often computationally expensive, recent studies mainly focus on decomposing it on spatial-temporal and/or channel dimensions. Unfortunately, most approaches fail to achieve a preferable balance between convolutional efficiency and feature-interaction sufficiency. For this reason, we propose a concise and novel Channel Tensorization Network (CT-Net), by treating the channel dimension of input feature as a multiplication of K sub-dimensions. On one hand, it naturally factorizes convolution in a multiple dimension way, leading to a light computation burden. On the other hand, it can effectively enhance feature interaction from different channels, and progressively enlarge the 3D receptive field of such interaction to boost classification accuracy. Furthermore, we equip our CT-Module with a Tensor Excitation (TE) mechanism. It can learn to exploit spatial, temporal and channel attention in a high-dimensional manner, to improve the cooperative power of all the feature dimensions in our CT-Module. Finally, we flexibly adapt ResNet as our CT-Net. Extensive experiments are conducted on several challenging video benchmarks, e.g., Kinetics-400, Something-Something V1 and V2. Our CT-Net outperforms a number of recent SOTA approaches, in terms of accuracy and/or efficiency.

Learning Invariant Representations for Reinforcement Learning without Reconstruction

- Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, Sergey Levine
- abstract@open-review(Oral): We study how representation learning can accelerate reinforcement learning from rich observations, such as images, without relying either on domain knowledge or pixel-reconstruction. Our goal is to learn representations that provide for effective downstream control and invariance to task-irrelevant details. Bisimulation metrics quantify behavioral similarity between states in continuous MDPs, which we propose using to learn robust latent representations which encode only the task-relevant information from observations. Our method trains encoders such that distances in latent space equal bisimulation distances in state space. We demonstrate the effectiveness of our method at disregarding task-irrelevant information using modified visual MuJoCo tasks, where the background is replaced with moving distractors and natural videos, while achieving SOTA performance. We also test a first-person highway driving task where our method learns invariance to clouds, weather, and time of day. Finally, we provide generalization results drawn from properties of bisimulation metrics, and links to causal inference.

Intraclass clustering: an implicit learning ability that regularizes DNNs

- Simon Carbonnelle, Christophe De Vleeschouwer
- abstract@open-review(Poster): Several works have shown that the regularization mechanisms underlying deep neural networks' generalization performances are still poorly understood. In this paper, we hypothesize that deep neural networks are regularized through their ability to extract meaningful clusters among the samples of a class. This constitutes an implicit form of regularization, as no explicit training mechanisms or supervision target such behaviour. To support our hypothesis, we design four different measures of intraclass clustering, based on the neuron- and layer-level representations of the training data. We then show that these measures constitute accurate predictors of generalization performance across variations of a large set of hyperparameters (learning rate, batch size, optimizer, weight decay, dropout rate, data augmentation, network depth and width).

Learning Robust State Abstractions for Hidden-Parameter Block MDPs

- Amy Zhang, Shagun Sodhani, Khimya Khetarpal, Joelle Pineau
- abstract@open-review(Poster): Many control tasks exhibit similar dynamics that can be modeled as having common latent structure. Hidden-Parameter Markov Decision Processes (HiP-MDPs) explicitly model this structure to improve sample efficiency in multi-task settings. However, this setting makes strong assumptions on the observability of the state that limit its application in real-world scenarios with rich observation spaces. In this work, we leverage ideas of common structure from the HiP-MDP setting, and extend it to enable robust state abstractions inspired by Block MDPs. We derive instantiations of this new framework for both multi-task reinforcement learning (MTRL) and meta-reinforcement learning (Meta-RL) settings. Further, we provide transfer and generalization bounds based on task and state similarity, along with sample complexity bounds that depend on the aggregate number of samples across tasks, rather than the number of tasks, a significant improvement over prior work. To further demonstrate efficacy of the proposed method, we empirically compare and show improvement over multi-task and meta-reinforcement learning baselines.

Isometric Transformation Invariant and Equivariant Graph Convolutional Networks

- Masanobu Horie, Naoki Morita, Toshiaki Hishinuma, Yu Ihara, Naoto Mitsume
- abstract@open-review(Poster): Graphs are one of the most important data structures for representing pairwise relations between objects. Specifically, a graph embedded in a Euclidean space is essential to solving real problems, such as physical simulations. A crucial requirement for applying graphs in Euclidean spaces to physical simulations is learning and inferring the isometric transformation invariant and equivariant features in a computationally efficient manner. In this paper, we propose a set of transformation invariant and equivariant models based on graph convolutional networks, called IsoGCNs. We demonstrate that the proposed model has a competitive performance compared to state-of-the-art methods on tasks related to geometrical and physical simulation data. Moreover, the proposed model can scale up to graphs with 1M vertices and conduct an inference faster than a conventional finite element analysis, which the existing equivariant models cannot achieve.

Learning Safe Multi-agent Control with Decentralized Neural Barrier Certificates

- Zengyi Qin, Kaiqing Zhang, Yuxiao Chen, Jingkai Chen, Chuchu Fan
- abstract@open-review(Poster): We study the multi-agent safe control problem where agents should avoid collisions to static obstacles and collisions with each other while reaching their goals. Our core idea is to learn the multi-agent control policy jointly with learning the control barrier functions as safety certificates. We propose a new joint-learning framework that can be implemented in a decentralized fashion, which can adapt to an arbitrarily large number of agents. Building upon this framework, we further improve the scalability by incorporating neural network architectures that are invariant to the quantity and permutation of neighboring agents. In addition, we propose a new spontaneous policy refinement method to further enforce the certificate condition during testing. We provide extensive experiments to demonstrate that our method significantly outperforms other leading multi-agent control approaches in terms of maintaining safety and completing original tasks. Our approach also shows substantial generalization capability in that the control policy can be trained with 8 agents in one scenario, while being used on other scenarios with up to 1024 agents in complex multi-agent environments and dynamics. Videos and source code can be found at https://realm.mit.edu/blog/learning-safe-multi-agent-control-decentralized-neural-barrier-certificates.

Learning Incompressible Fluid Dynamics from Scratch - Towards Fast, Differentiable Fluid Models that Generalize

- Nils Wandel, Michael Weinmann, Reinhard Klein
- abstract@open-review(Spotlight): Fast and stable fluid simulations are an essential prerequisite for applications ranging from computer-generated imagery to computer-aided design in research and development. However, solving the partial differential equations of incompressible fluids is a challenging task and traditional numerical approximation schemes come at high computational costs. Recent deep learning based approaches promise vast speed-ups but do not generalize to new fluid domains, require fluid simulation data for training, or rely on complex pipelines that outsource major parts of the fluid simulation to traditional methods.

In this work, we propose a novel physics-constrained training approach that generalizes to new fluid domains, requires no fluid simulation data, and allows convolutional neural networks to map a fluid state from time-point t to a subsequent state at time t+dt in a single forward pass. This simplifies the pipeline to train and evaluate neural fluid models. After training, the framework yields models that are capable of fast fluid simulations and can handle various fluid phenomena including the Magnus effect and Kármán vortex streets. We present an interactive real-time demo to show the speed and generalization capabilities of our trained models. Moreover, the trained neural networks are efficient differentiable fluid solvers as they offer a differentiable update step to advance the fluid simulation in time. We exploit this fact in a proof-of-concept optimal control experiment. Our models significantly outperform a recent differentiable fluid solver in terms of computational speed and accuracy.

ChipNet: Budget-Aware Pruning with Heaviside Continuous Approximations

- Rishabh Tiwari, Udbhav Bamba, Arnav Chavan, Deepak Gupta
- abstract@open-review(Poster): Structured pruning methods are among the effective strategies for extracting small resource-efficient convolutional neural networks from their dense counterparts with minimal loss in accuracy. However, most existing methods still suffer from one or more limitations, that include 1) the need for training the dense model from scratch with pruning-related parameters embedded in the architecture, 2) requiring model-specific hyperparameter settings, 3) inability to include budget-related constraint in the training process, and 4) instability under scenarios of extreme pruning. In this paper, we present ChipNet, a deterministic pruning strategy that employs continuous Heaviside function and a novel crispness loss to identify a highly sparse network out of an existing dense network. Our choice of continuous Heaviside function is inspired by the field of design optimization, where the material distribution task is posed as a continuous optimization problem, but only discrete values (0 or 1) are practically feasible and expected as final outcomes. Our approach's flexible design facilitates its use with different choices of budget constraints while maintaining stability for very low target budgets. Experimental results show that ChipNet outperforms state-of-the-art structured pruning methods by remarkable margins of up to 16.1% in terms of accuracy. Further, we show that the masks obtained with ChipNet are transferable across datasets. For certain cases, it was observed that masks transferred from a model trained on feature-rich teacher dataset provide better performance on the student dataset than those obtained by directly pruning on the student data itself.

AdaSpeech: Adaptive Text to Speech for Custom Voice

- Mingjian Chen, Xu Tan, Bohan Li, Yanqing Liu, Tao Qin, sheng zhao, Tie-Yan Liu
- abstract@open-review(Poster): Custom voice, a specific text to speech (TTS) service in commercial speech platforms, aims to adapt a source TTS model to synthesize personal voice for a target speaker using few speech from her/him. Custom voice presents two unique challenges for TTS adaptation: 1) to support diverse customers, the adaptation model needs to handle diverse acoustic conditions which could be very different from source speech data, and 2) to support a large number of customers, the adaptation parameters need to be small enough for each target speaker to reduce memory usage while maintaining high voice quality. In this work, we propose AdaSpeech, an adaptive TTS system for high-quality and efficient customization of new voices. We design several techniques in AdaSpeech to address the two challenges in custom voice: 1) To handle different acoustic conditions, we model the acoustic information in both utterance and phoneme level. Specifically, we use one acoustic encoder to extract an utterance-level vector and another one to extract a sequence of phoneme-level vectors from the target speech during pre-training and fine-tuning; in inference, we extract the utterance-level vector from a reference speech and use an acoustic predictor to predict the phoneme-level vectors. 2) To better trade off the adaptation parameters and voice quality, we introduce conditional layer normalization in the mel-spectrogram decoder of AdaSpeech, and fine-tune this part in addition to speaker embedding for adaptation. We pre-train the source TTS model on LibriTTS datasets and fine-tune it on VCTK and LJSpeech datasets (with different acoustic conditions from LibriTTS) with few adaptation data, e.g., 20 sentences, about 1 minute speech. Experiment results show that AdaSpeech achieves much better adaptation quality than baseline methods, with only about 5K specific parameters for each speaker, which demonstrates its effectiveness for custom voice. The audio samples are available at https://speechresearch.github.io/adaspeech/.

Estimating and Evaluating Regression Predictive Uncertainty in Deep Object Detectors

- Ali Harakeh, Steven L. Waslander
- abstract@open-review(Poster): Predictive uncertainty estimation is an essential next step for the reliable deployment of deep object detectors in safety-critical tasks. In this work, we focus on estimating predictive distributions for bounding box regression output with variance networks. We show that in the context of object detection, training variance networks with negative log likelihood (NLL) can lead to high entropy predictive distributions regardless of the correctness of the output mean. We propose to use the energy score as a non-local proper scoring rule and find that when used for training, the energy score leads to better calibrated and lower entropy predictive distributions than NLL. We also address the widespread use of non-proper scoring metrics for evaluating predictive distributions from deep object detectors by proposing an alternate evaluation approach founded on proper scoring rules. Using the proposed evaluation tools, we show that although variance networks can be used to produce high quality predictive distributions, ad-hoc approaches used by seminal object detectors for choosing regression targets during training do not provide wide enough data support for reliable variance learning. We hope that our work helps shift evaluation in probabilistic object detection to better align with predictive uncertainty evaluation in other machine learning domains. Code for all models, evaluation, and datasets is available at: https://github.com/asharakeh/probdet.git.

Domain-Robust Visual Imitation Learning with Mutual Information Constraints

- Edoardo Cetin, Oya Celiktutan
- abstract@open-review(Poster): Human beings are able to understand objectives and learn by simply observing others perform a task. Imitation learning methods aim to replicate such capabilities, however, they generally depend on access to a full set of optimal states and actions taken with the agent's actuators and from the agent's point of view. In this paper, we introduce a new algorithm - called Disentangling Generative Adversarial Imitation Learning (DisentanGAIL) - with the purpose of bypassing such constraints. Our algorithm enables autonomous agents to learn directly from high dimensional observations of an expert performing a task, by making use of adversarial learning with a latent representation inside the discriminator network. Such latent representation is regularized through mutual information constraints to incentivize learning only features that encode information about the completion levels of the task being demonstrated. This allows to obtain a shared feature space to successfully perform imitation while disregarding the differences between the expert's and the agent's domains. Empirically, our algorithm is able to efficiently imitate in a diverse range of control problems including balancing, manipulation and locomotive tasks, while being robust to various domain differences in terms of both environment appearance and agent embodiment.

Clustering-friendly Representation Learning via Instance Discrimination and Feature Decorrelation

- Yaling Tao, Kentaro Takagi, Kouta Nakata
- abstract@open-review(Poster): Clustering is one of the most fundamental tasks in machine learning. Recently, deep clustering has become a major trend in clustering techniques. Representation learning often plays an important role in the effectiveness of deep clustering, and thus can be a principal cause of performance degradation. In this paper, we propose a clustering-friendly representation learning method using instance discrimination and feature decorrelation. Our deep-learning-based representation learning method is motivated by the properties of classical spectral clustering. Instance discrimination learns similarities among data and feature decorrelation removes redundant correlation among features. We utilize an instance discrimination method in which learning individual instance classes leads to learning similarity among instances. Through detailed experiments and examination, we show that the approach can be adapted to learning a latent space for clustering. We design novel softmax-formulated decorrelation constraints for learning. In evaluations of image clustering using CIFAR-10 and ImageNet-10, our method achieves accuracy of 81.5% and 95.4%, respectively. We also show that the softmax-formulated constraints are compatible with various neural networks.

A Gradient Flow Framework For Analyzing Network Pruning

- Ekdeep Singh Lubana, Robert Dick
- abstract@open-review(Spotlight): Recent network pruning methods focus on pruning models early-on in training. To estimate the impact of removing a parameter, these methods use importance measures that were originally designed to prune trained models. Despite lacking justification for their use early-on in training, such measures result in surprisingly low accuracy loss. To better explain this behavior, we develop a general framework that uses gradient flow to unify state-of-the-art importance measures through the norm of model parameters. We use this framework to determine the relationship between pruning measures and evolution of model parameters, establishing several results related to pruning models early-on in training: (i) magnitude-based pruning removes parameters that contribute least to reduction in loss, resulting in models that converge faster than magnitude-agnostic methods; (ii) loss-preservation based pruning preserves first-order model evolution dynamics and its use is therefore justified for pruning minimally trained models; and (iii) gradient-norm based pruning affects second-order model evolution dynamics, such that increasing gradient norm via pruning can produce poorly performing models. We validate our claims on several VGG-13, MobileNet-V1, and ResNet-56 models trained on CIFAR-10/CIFAR-100.

Progressive Skeletonization: Trimming more fat from a network at initialization

- Pau de Jorge, Amartya Sanyal, Harkirat Behl, Philip Torr, Grégory Rogez, Puneet K. Dokania
- abstract@open-review(Poster): Recent studies have shown that skeletonization (pruning parameters) of networks at initialization provides all the practical benefits of sparsity both at inference and training time, while only marginally degrading their performance. However, we observe that beyond a certain level of sparsity (approx 95%), these approaches fail to preserve the network performance, and to our surprise, in many cases perform even worse than trivial random pruning. To this end, we propose an objective to find a skeletonized network with maximum foresight connection sensitivity (FORCE) whereby the trainability, in terms of connection sensitivity, of a pruned network is taken into consideration. We then propose two approximate procedures to maximize our objective (1) Iterative SNIP: allows parameters that were unimportant at earlier stages of skeletonization to become important at later stages; and (2) FORCE: iterative process that allows exploration by allowing already pruned parameters to resurrect at later stages of skeletonization. Empirical analysis on a large suite of experiments show that our approach, while providing at least as good performance as other recent approaches on moderate pruning levels, provide remarkably improved performance on high pruning levels (could remove up to 99.5% parameters while keeping the networks trainable).

On the Curse of Memory in Recurrent Neural Networks: Approximation and Optimization Analysis

- Zhong Li, Jiequn Han, Weinan E, Qianxiao Li
- abstract@open-review(Poster): We study the approximation properties and optimization dynamics of recurrent neural networks (RNNs) when applied to learn input-output relationships in temporal data. We consider the simple but representative setting of using continuous-time linear RNNs to learn from data generated by linear relationships. Mathematically, the latter can be understood as a sequence of linear functionals. We prove a universal approximation theorem of such linear functionals and characterize the approximation rate. Moreover, we perform a fine-grained dynamical analysis of training linear RNNs by gradient methods. A unifying theme uncovered is the non-trivial effect of memory, a notion that can be made precise in our framework, on both approximation and optimization: when there is long-term memory in the target, it takes a large number of neurons to approximate it. Moreover, the training process will suffer from slow downs. In particular, both of these effects become exponentially more pronounced with increasing memory - a phenomenon we call the “curse of memory”. These analyses represent a basic step towards a concrete mathematical understanding of new phenomenons that may arise in learning temporal relationships using recurrent architectures.

Do not Let Privacy Overbill Utility: Gradient Embedding Perturbation for Private Learning

- Da Yu, Huishuai Zhang, Wei Chen, Tie-Yan Liu
- abstract@open-review(Poster): The privacy leakage of the model about the training data can be bounded in the differential privacy mechanism. However, for meaningful privacy parameters, a differentially private model degrades the utility drastically when the model comprises a large number of trainable parameters. In this paper, we propose an algorithm \emph{Gradient Embedding Perturbation (GEP)} towards training differentially private deep models with decent accuracy. Specifically, in each gradient descent step, GEP first projects individual private gradient into a non-sensitive anchor subspace, producing a low-dimensional gradient embedding and a small-norm residual gradient. Then, GEP perturbs the low-dimensional embedding and the residual gradient separately according to the privacy budget. Such a decomposition permits a small perturbation variance, which greatly helps to break the dimensional barrier of private learning. With GEP, we achieve decent accuracy with low computational cost and modest privacy guarantee for deep models. Especially, with privacy bound $\epsilon=8$, we achieve 74.9% test accuracy on CIFAR10 and 95.1% test accuracy on SVHN, significantly improving over existing results.

More or Less: When and How to Build Convolutional Neural Network Ensembles

- Abdul Wasay, Stratos Idreos
- abstract@open-review(Poster): Convolutional neural networks are utilized to solve increasingly more complex problems and with more data. As a result, researchers and practitioners seek to scale the representational power of such models by adding more parameters. However, increasing parameters requires additional critical resources in terms of memory and compute, leading to increased training and inference cost. Thus a consistent challenge is to obtain as high as possible accuracy within a parameter budget. As neural network designers navigate this complex landscape, they are guided by conventional wisdom that is informed from past empirical studies. We identify a critical part of this design space that is not well-understood: How to decide between the alternatives of expanding a single convolutional network model or increasing the number of networks in the form of an ensemble. We study this question in detail across various network architectures and data sets. We build an extensive experimental framework that captures numerous angles of the possible design space in terms of how a new set of parameters can be used in a model. We consider a holistic set of metrics such as training time, inference time, and memory usage. The framework provides a robust assessment by making sure it controls for the number of parameters. Contrary to conventional wisdom, we show that when we perform a holistic and robust assessment, we uncover a wide design space, where ensembles provide better accuracy, train faster, and deploy at speed comparable to single convolutional networks with the same total number of parameters.

Towards Robustness Against Natural Language Word Substitutions

- Xinshuai Dong, Anh Tuan Luu, Rongrong Ji, Hong Liu
- abstract@open-review(Spotlight): Robustness against word substitutions has a well-defined and widely acceptable form, i.e., using semantically similar words as substitutions, and thus it is considered as a fundamental stepping-stone towards broader robustness in natural language processing. Previous defense methods capture word substitutions in vector space by using either l_2-ball or hyper-rectangle, which results in perturbation sets that are not inclusive enough or unnecessarily large, and thus impedes mimicry of worst cases for robust training. In this paper, we introduce a novel Adversarial Sparse Convex Combination (ASCC) method. We model the word substitution attack space as a convex hull and leverages a regularization term to enforce perturbation towards an actual substitution, thus aligning our modeling better with the discrete textual space. Based on ASCC method, we further propose ASCC-defense, which leverages ASCC to generate worst-case perturbations and incorporates adversarial training towards robustness. Experiments show that ASCC-defense outperforms the current state-of-the-arts in terms of robustness on two prevailing NLP tasks, i.e., sentiment analysis and natural language inference, concerning several attacks across multiple model architectures. Besides, we also envision a new class of defense towards robustness in NLP, where our robustly trained word vectors can be plugged into a normally trained model and enforce its robustness without applying any other defense techniques.

Revisiting Hierarchical Approach for Persistent Long-Term Video Prediction

- Wonkwang Lee, Whie Jung, Han Zhang, Ting Chen, Jing Yu Koh, Thomas Huang, Hyungsuk Yoon, Honglak Lee, Seunghoon Hong
- abstract@open-review(Poster): Learning to predict the long-term future of video frames is notoriously challenging due to the inherent ambiguities in a distant future and dramatic amplification of prediction error over time. Despite the recent advances in the literature, existing approaches are limited to moderately short-term prediction (less than a few seconds), while extrapolating it to a longer future quickly leads to destruction in structure and content. In this work, we revisit the hierarchical models in video prediction. Our method generates future frames by first estimating a sequence of dense semantic structures and subsequently translating the estimated structures to pixels by video-to-video translation model. Despite the simplicity, we show that modeling structures and their dynamics in categorical structure space with stochastic sequential estimator leads to surprisingly successful long-term prediction. We evaluate our method on two challenging video prediction scenarios, \emph{car driving} and \emph{human dancing}, and demonstrate that it can generate complicated scene structures and motions over a very long time horizon (\ie~thousands frames), setting a new standard of video prediction with orders of magnitude longer prediction time than existing approaches. Video results are available at https://1konny.github.io/HVP/.

Symmetry-Aware Actor-Critic for 3D Molecular Design

- Gregor N. C. Simm, Robert Pinsler, Gábor Csányi, José Miguel Hernández-Lobato
- abstract@open-review(Poster): Automating molecular design using deep reinforcement learning (RL) has the potential to greatly accelerate the search for novel materials. Despite recent progress on leveraging graph representations to design molecules, such methods are fundamentally limited by the lack of three-dimensional (3D) information. In light of this, we propose a novel actor-critic architecture for 3D molecular design that can generate molecular structures unattainable with previous approaches. This is achieved by exploiting the symmetries of the design process through a rotationally covariant state-action representation based on a spherical harmonics series expansion. We demonstrate the benefits of our approach on several 3D molecular design tasks, where we find that building in such symmetries significantly improves generalization and the quality of generated molecules.

Learnable Embedding sizes for Recommender Systems

- Siyi Liu, Chen Gao, Yihong Chen, Depeng Jin, Yong Li
- abstract@open-review(Poster): The embedding-based representation learning is commonly used in deep learning recommendation models to map the raw sparse features to dense vectors. The traditional embedding manner that assigns a uniform size to all features has two issues. First, the numerous features inevitably lead to a gigantic embedding table that causes a high memory usage cost. Second, it is likely to cause the over-fitting problem for those features that do not require too large representation capacity. Existing works that try to address the problem always cause a significant drop in recommendation performance or suffers from the limitation of unaffordable training time cost. In this paper, we proposed a novel approach, named PEP (short for Plug-in Embedding Pruning), to reduce the size of the embedding table while avoiding the drop of recommendation accuracy. PEP prunes embedding parameter where the pruning threshold(s) can be adaptively learned from data. Therefore we can automatically obtain a mixed-dimension embedding-scheme by pruning redundant parameters for each feature. PEP is a general framework that can plug in various base recommendation models. Extensive experiments demonstrate it can efficiently cut down embedding parameters and boost the base model's performance. Specifically, it achieves strong recommendation performance while reducing 97-99% parameters. As for the computation cost, PEP only brings an additional 20-30% time cost compare with base models.

Iterative Empirical Game Solving via Single Policy Best Response

- Max Smith, Thomas Anthony, Michael Wellman
- abstract@open-review(Spotlight): Policy-Space Response Oracles (PSRO) is a general algorithmic framework for learning policies in multiagent systems by interleaving empirical game analysis with deep reinforcement learning (DRL). At each iteration, DRL is invoked to train a best response to a mixture of opponent policies. The repeated application of DRL poses an expensive computational burden as we look to apply this algorithm to more complex domains. We introduce two variations of PSRO designed to reduce the amount of simulation required during DRL training. Both algorithms modify how PSRO adds new policies to the empirical game, based on learned responses to a single opponent policy. The first, Mixed-Oracles, transfers knowledge from previous iterations of DRL, requiring training only against the opponent's newest policy. The second, Mixed-Opponents, constructs a pure-strategy opponent by mixing existing strategy's action-value estimates, instead of their policies. Learning against a single policy mitigates conflicting experiences on behalf of a learner facing an unobserved distribution of opponents. We empirically demonstrate that these algorithms substantially reduce the amount of simulation during training required by PSRO, while producing equivalent or better solutions to the game.

Learning Neural Generative Dynamics for Molecular Conformation Generation

- Minkai Xu, Shitong Luo, Yoshua Bengio, Jian Peng, Jian Tang
- abstract@open-review(Poster): We study how to generate molecule conformations (i.e., 3D structures) from a molecular graph. Traditional methods, such as molecular dynamics, sample conformations via computationally expensive simulations. Recently, machine learning methods have shown great potential by training on a large collection of conformation data. Challenges arise from the limited model capacity for capturing complex distributions of conformations and the difficulty in modeling long-range dependencies between atoms. Inspired by the recent progress in deep generative models, in this paper, we propose a novel probabilistic framework to generate valid and diverse conformations given a molecular graph. We propose a method combining the advantages of both flow-based and energy-based models, enjoying: (1) a high model capacity to estimate the multimodal conformation distribution; (2) explicitly capturing the complex long-range dependencies between atoms in the observation space. Extensive experiments demonstrate the superior performance of the proposed method on several benchmarks, including conformation generation and distance modeling tasks, with a significant improvement over existing generative models for molecular conformation sampling.

Learning the Pareto Front with Hypernetworks

- Aviv Navon, Aviv Shamsian, Ethan Fetaya, Gal Chechik
- abstract@open-review(Poster): Multi-objective optimization (MOO) problems are prevalent in machine learning. These problems have a set of optimal solutions, called the Pareto front, where each point on the front represents a different trade-off between possibly conflicting objectives. Recent MOO methods can target a specific desired ray in loss space however, most approaches still face two grave limitations: (i) A separate model has to be trained for each point on the front; and (ii) The exact trade-off must be known before the optimization process. Here, we tackle the problem of learning the entire Pareto front, with the capability of selecting a desired operating point on the front after training. We call this new setup Pareto-Front Learning (PFL).

We describe an approach to PFL implemented using HyperNetworks, which we term Pareto HyperNetworks (PHNs). PHN learns the entire Pareto front simultaneously using a single hypernetwork, which receives as input a desired preference vector and returns a Pareto-optimal model whose loss vector is in the desired ray. The unified model is runtime efficient compared to training multiple models and generalizes to new operating points not used during training. We evaluate our method on a wide set of problems, from multi-task regression and classification to fairness. PHNs learn the entire Pareto front at roughly the same time as learning a single point on the front and at the same time reach a better solution set. PFL opens the door to new applications where models are selected based on preferences that are only available at run time.

Ask Your Humans: Using Human Instructions to Improve Generalization in Reinforcement Learning

- Valerie Chen, Abhinav Gupta, Kenneth Marino

- abstract@open-review(Poster): Complex, multi-task problems have proven to be difficult to solve efficiently in a sparse-reward reinforcement learning setting. In order to be sample efficient, multi-task learning requires reuse and sharing of low-level policies. To facilitate the automatic decomposition of hierarchical tasks, we propose the use of step-by-step human demonstrations in the form of natural language instructions and action trajectories. We introduce a dataset of such demonstrations in a crafting-based grid world. Our model consists of a high-level language generator and low-level policy, conditioned on language. We find that human demonstrations help solve the most complex tasks. We also find that incorporating natural language allows the model to generalize to unseen tasks in a zero-shot setting and to learn quickly from a few demonstrations. Generalization is not only reflected in the actions of the agent, but also in the generated natural language instructions in unseen tasks. Our approach also gives our trained agent interpretable behaviors because it is able to generate a sequence of high-level descriptions of its actions.

Taming GANs with Lookahead-Minmax

- Tatjana Chavdarova, Matteo Pagliardini, Sebastian U Stich, François Fleuret, Martin Jaggi
- abstract@open-review(Poster): Generative Adversarial Networks are notoriously challenging to train. The underlying minmax optimization is highly susceptible to the variance of the stochastic gradient and the rotational component of the associated game vector field. To tackle these challenges, we propose the Lookahead algorithm for minmax optimization, originally developed for single objective minimization only. The backtracking step of our Lookahead–minmax naturally handles the rotational game dynamics, a property which was identified to be key for enabling gradient ascent descent methods to converge on challenging examples often analyzed in the literature. Moreover, it implicitly handles high variance without using large mini-batches, known to be essential for reaching state of the art performance. Experimental results on MNIST, SVHN, CIFAR-10, and ImageNet demonstrate a clear advantage of combining Lookahead–minmax with Adam or extragradient, in terms of performance and improved stability, for negligible memory and computational cost. Using 30-fold fewer parameters and 16-fold smaller minibatches we outperform the reported performance of the class-dependent BigGAN on CIFAR-10 by obtaining FID of 12.19 without using the class labels, bringing state-of-the-art GAN training within reach of common computational resources.

Uncertainty in Gradient Boosting via Ensembles

- Andrey Malinin, Liudmila Prokhorenkova, Aleksei Ustimenko
- abstract@open-review(Poster): For many practical, high-risk applications, it is essential to quantify uncertainty in a model's predictions to avoid costly mistakes. While predictive uncertainty is widely studied for neural networks, the topic seems to be under-explored for models based on gradient boosting. However, gradient boosting often achieves state-of-the-art results on tabular data. This work examines a probabilistic ensemble-based framework for deriving uncertainty estimates in the predictions of gradient boosting classification and regression models. We conducted experiments on a range of synthetic and real datasets and investigated the applicability of ensemble approaches to gradient boosting models that are themselves ensembles of decision trees. Our analysis shows that ensembles of gradient boosting models successfully detect anomalous inputs while having limited ability to improve the predicted total uncertainty. Importantly, we also propose a concept of a virtual ensemble to get the benefits of an ensemble via only one gradient boosting model, which significantly reduces complexity.

Adversarial score matching and improved sampling for image generation

- Alexia Jolicoeur-Martineau, Rémi Piché-Taillefer, Ioannis Mitliagkas, Remi Tachet des Combes
- abstract@open-review(Poster): Denoising Score Matching with Annealed Langevin Sampling (DSM-ALS) has recently found success in generative modeling. The approach works by first training a neural network to estimate the score of a distribution, and then using Langevin dynamics to sample from the data distribution assumed by the score network. Despite the convincing visual quality of samples, this method appears to perform worse than Generative Adversarial Networks (GANs) under the Fréchet Inception Distance, a standard metric for generative models. We show that this apparent gap vanishes when denoising the final Langevin samples using the score network. In addition, we propose two improvements to DSM-ALS: 1) Consistent Annealed Sampling as a more stable alternative to Annealed Langevin Sampling, and 2) a hybrid training formulation, composed of both Denoising Score Matching and adversarial objectives. By combining these two techniques and exploring different network architectures, we elevate score matching methods and obtain results competitive with state-of-the-art image generation on CIFAR-10.

FastSpeech 2: Fast and High-Quality End-to-End Text to Speech

- Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, Tie-Yan Liu
- abstract@open-review(Poster): Non-autoregressive text to speech (TTS) models such as FastSpeech can synthesize speech significantly faster than previous autoregressive models with comparable quality. The training of FastSpeech model relies on an autoregressive teacher model for duration prediction (to provide more information as input) and knowledge distillation (to simplify the data distribution in output), which can ease the one-to-many mapping problem (i.e., multiple speech variations correspond to the same text) in TTS. However, FastSpeech has several disadvantages: 1) the teacher-student distillation pipeline is complicated and time-consuming, 2) the duration extracted from the teacher model is not accurate enough, and the target mel-spectrograms distilled from teacher model suffer from information loss due to data simplification, both of which limit the voice quality. In this paper, we propose FastSpeech 2, which addresses the issues in FastSpeech and better solves the one-to-many mapping problem in TTS by 1) directly training the model with ground-truth target instead of the simplified output from teacher, and 2) introducing more variation information of speech (e.g., pitch, energy and more accurate duration) as conditional inputs. Specifically, we extract duration, pitch and energy from speech waveform and directly take them as conditional inputs in training and use predicted values in inference. We further design FastSpeech 2s, which is the first attempt to directly generate speech waveform from text in parallel, enjoying the benefit of fully end-to-end inference. Experimental results show that 1) FastSpeech 2 achieves a 3x training speed-up over FastSpeech, and FastSpeech 2s enjoys even faster inference speed; 2) FastSpeech 2 and 2s outperform FastSpeech in voice quality, and FastSpeech 2 can even surpass autoregressive models. Audio samples are available at https://speechresearch.github.io/fastspeech2/.

Interpretable Models for Granger Causality Using Self-explaining Neural Networks

- Ričards Marcinkevičs, Julia E Vogt
- abstract@open-review(Poster): Exploratory analysis of time series data can yield a better understanding of complex dynamical systems. Granger causality is a practical framework for analysing interactions in sequential data, applied in a wide range of domains. In this paper, we propose a novel framework for inferring multivariate Granger causality under nonlinear dynamics based on an extension of self-explaining neural networks. This framework is more interpretable than other neural-network-based techniques for inferring Granger causality, since in addition to relational inference, it also allows detecting signs of Granger-causal effects and inspecting their variability over time. In comprehensive experiments on simulated data, we show that our framework performs on par with several powerful baseline methods at inferring Granger causality and that it achieves better performance at inferring interaction signs. The results suggest that our framework is a viable and more interpretable alternative to sparse-input neural networks for inferring Granger causality.

Learning Deep Features in Instrumental Variable Regression

- Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, Arthur Gretton
- abstract@open-review(Poster): Instrumental variable (IV) regression is a standard strategy for learning causal relationships between confounded treatment and outcome variables from observational data by using an instrumental variable, which affects the outcome only through the treatment. In classical IV regression, learning proceeds in two stages: stage 1 performs linear regression from the instrument to the treatment; and stage 2 performs linear regression from the treatment to the outcome, conditioned on the instrument. We propose a novel method, deep feature instrumental variable regression (DFIV), to

address the case where relations between instruments, treatments, and outcomes may be nonlinear. In this case, deep neural nets are trained to define informative nonlinear features on the instruments and treatments. We propose an alternating training regime for these features to ensure good end-to-end performance when composing stages 1 and 2, thus obtaining highly flexible feature maps in a computationally efficient manner. DFIV outperforms recent state-of-the-art methods on challenging IV benchmarks, including settings involving high dimensional image data. DFIV also exhibits competitive performance in off-policy policy evaluation for reinforcement learning, which can be understood as an IV regression task.

Statistical inference for individual fairness

- Subha Maity, Songkai Xue, Mikhail Yurochkin, Yuekai Sun
- abstract@open-review(Poster): As we rely on machine learning (ML) models to make more consequential decisions, the issue of ML models perpetuating unwanted social biases has come to the fore of the public's and the research community's attention. In this paper, we focus on the problem of detecting violations of individual fairness in ML models. We formalize the problem as measuring the susceptibility of ML models against a form of adversarial attack and develop a suite of inference tools for the adversarial loss. The tools allow practitioners to assess the individual fairness of ML models in a statistically-principled way: form confidence intervals for the adversarial loss and test hypotheses of model fairness with (asymptotic) non-coverage/Type I error rate control. We demonstrate the utility of our tools in a real-world case study.

Self-Supervised Policy Adaptation during Deployment

- Nicklas Hansen, Rishabh Jangir, Yu Sun, Guillem Alenyà, Pieter Abbeel, Alexei A Efros, Lerrel Pinto, Xiaolong Wang
- abstract@open-review(Spotlight): In most real world scenarios, a policy trained by reinforcement learning in one environment needs to be deployed in another, potentially quite different environment. However, generalization across different environments is known to be hard. A natural solution would be to keep training after deployment in the new environment, but this cannot be done if the new environment offers no reward signal. Our work explores the use of self-supervision to allow the policy to continue training after deployment without using any rewards. While previous methods explicitly anticipate changes in the new environment, we assume no prior knowledge of those changes yet still obtain significant improvements. Empirical evaluations are performed on diverse simulation environments from DeepMind Control suite and ViZDoom, as well as real robotic manipulation tasks in continuously changing environments, taking observations from an uncalibrated camera. Our method improves generalization in 31 out of 36 environments across various tasks and outperforms domain randomization on a majority of environments. Webpage and implementation: https://nicklashansen.github.io/PAD/.

Randomized Ensembled Double Q-Learning: Learning Fast Without a Model

- Xinyue Chen, Che Wang, Zijian Zhou, Keith W. Ross
- abstract@open-review(Poster): Using a high Update-To-Data (UTD) ratio, model-based methods have recently achieved much higher sample efficiency than previous model-free methods for continuous-action DRL benchmarks. In this paper, we introduce a simple model-free algorithm, Randomized Ensembled Double Q-Learning (REDQ), and show that its performance is just as good as, if not better than, a state-of-the-art model-based algorithm for the MuJoCo benchmark. Moreover, REDQ can achieve this performance using fewer parameters than the model-based method, and with less wall-clock run time. REDQ has three carefully integrated ingredients which allow it to achieve its high performance: (i) a UTD ratio $\gg 1$; (ii) an ensemble of Q functions; (iii) in-target minimization across a random subset of Q functions from the ensemble. Through carefully designed experiments, we provide a detailed analysis of REDQ and related model-free algorithms. To our knowledge, REDQ is the first successful model-free DRL algorithm for continuous-action spaces using a UTD ratio $\gg 1$.

Lifelong Learning of Compositional Structures

- Jorge A Mendez, ERIC EATON
- abstract@open-review(Poster): A hallmark of human intelligence is the ability to construct self-contained chunks of knowledge and adequately reuse them in novel combinations for solving different yet structurally related problems. Learning such compositional structures has been a significant challenge for artificial systems, due to the combinatorial nature of the underlying search problem. To date, research into compositional learning has largely proceeded separately from work on lifelong or continual learning. We integrate these two lines of work to present a general-purpose framework for lifelong learning of compositional structures that can be used for solving a stream of related tasks. Our framework separates the learning process into two broad stages: learning how to best combine existing components in order to assimilate a novel problem, and learning how to adapt the set of existing components to accommodate the new problem. This separation explicitly handles the trade-off between the stability required to remember how to solve earlier tasks and the flexibility required to solve new tasks, as we show empirically in an extensive evaluation.

QPLEX: Duplex Dueling Multi-Agent Q-Learning

- Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, Chongjie Zhang
- abstract@open-review(Poster): We explore value-based multi-agent reinforcement learning (MARL) in the popular paradigm of centralized training with decentralized execution (CTDE). CTDE has an important concept, Individual-Global-Max (IGM) principle, which requires the consistency between joint and local action selections to support efficient local decision-making. However, in order to achieve scalability, existing MARL methods either limit representation expressiveness of their value function classes or relax the IGM consistency, which may suffer from instability risk or may not perform well in complex domains. This paper presents a novel MARL approach, called duPLEX dueling multi-agent Q-learning (QPLEX), which takes a duplex dueling network architecture to factorize the joint value function. This duplex dueling structure encodes the IGM principle into the neural network architecture and thus enables efficient value function learning. Theoretical analysis shows that QPLEX achieves a complete IGM function class. Empirical experiments on StarCraft II micromanagement tasks demonstrate that QPLEX significantly outperforms state-of-the-art baselines in both online and offline data collection settings, and also reveal that QPLEX achieves high sample efficiency and can benefit from offline datasets without additional online exploration.

Meta-Learning of Structured Task Distributions in Humans and Machines

- Sreejan Kumar, Ishita Dasgupta, Jonathan Cohen, Nathaniel Daw, Thomas Griffiths
- abstract@open-review(Poster): In recent years, meta-learning, in which a model is trained on a family of tasks (i.e. a task distribution), has emerged as an approach to training neural networks to perform tasks that were previously assumed to require structured representations, making strides toward closing the gap between humans and machines. However, we argue that evaluating meta-learning remains a challenge, and can miss whether meta-learning actually uses the structure embedded within the tasks. These meta-learners might therefore still be significantly different from humans learners. To demonstrate this difference, we first define a new meta-reinforcement learning task in which a structured task distribution is generated using a compositional grammar. We then introduce a novel approach to constructing a "null task distribution" with the same statistical complexity as this structured task distribution but without the explicit rule-based structure used to generate the structured task. We train a standard meta-learning agent, a recurrent network trained with model-free reinforcement learning, and compare it with human performance across the two task distributions. We find a double dissociation in which humans do better in the structured task distribution whereas agents do better in the null task distribution -- despite comparable statistical complexity. This work highlights that multiple strategies can achieve reasonable meta-test performance, and that careful construction of control task distributions is a valuable way to understand which strategies meta-learners acquire, and how they might differ from humans.

Differentially Private Learning Needs Better Features (or Much More Data)

- Florian Tramer, Dan Boneh
- abstract@open-review(Spotlight): We demonstrate that differentially private machine learning has not yet reached its "AlexNet moment" on many canonical vision tasks: linear models trained on handcrafted features significantly outperform end-to-end deep neural networks for moderate privacy budgets. To exceed the performance of handcrafted features, we show that private learning requires either much more private data, or access to features learned on public data from a similar domain. Our work introduces simple yet strong baselines for differentially private learning that can inform the evaluation of future progress in this area.

PC2WF: 3D Wireframe Reconstruction from Raw Point Clouds

- Yujia Liu, Stefano D'Aronco, Konrad Schindler, Jan Dirk Wegner
- abstract@open-review(Poster): We introduce PC2WF, the first end-to-end trainable deep network architecture to convert a 3D point cloud into a wireframe model. The network takes as input an unordered set of 3D points sampled from the surface of some object, and outputs a wireframe of that object, i.e., a sparse set of corner points linked by line segments. Recovering the wireframe is a challenging task, where the numbers of both vertices and edges are different for every instance, and a-priori unknown. Our architecture gradually builds up the model: It starts by encoding the points into feature vectors. Based on those features, it identifies a pool of candidate vertices, then prunes those candidates to a final set of corner vertices and refines their locations. Next, the corners are linked with an exhaustive set of candidate edges, which is again pruned to obtain the final wireframe. All steps are trainable, and errors can be backpropagated through the entire sequence. We validate the proposed model on a publicly available synthetic dataset, for which the ground truth wireframes are accessible, as well as on a new real-world dataset. Our model produces wireframe abstractions of good quality and outperforms several baselines.

Rethinking the Role of Gradient-based Attribution Methods for Model Interpretability

- Suraj Srinivas, Francois Fleuret
- abstract@open-review(Oral): Current methods for the interpretability of discriminative deep neural networks commonly rely on the model's input-gradients, i.e., the gradients of the output logits w.r.t. the inputs. The common assumption is that these input-gradients contain information regarding $p_{\theta} (y\mid \mathbf{x})$, the model's discriminative capabilities, thus justifying their use for interpretability. However, in this work, we show that these input-gradients can be arbitrarily manipulated as a consequence of the shift-invariance of softmax without changing the discriminative function. This leaves an open question: given that input-gradients can be arbitrary, why are they highly structured and explanatory in standard models?

In this work, we re-interpret the logits of standard softmax-based classifiers as unnormalized log-densities of the data distribution and show that input-gradients can be viewed as gradients of a class-conditional generative model $p_{\theta}(\mathbf{x} \mid y)$ implicit in the discriminative model. This leads us to hypothesize that the highly structured and explanatory nature of input-gradients may be due to the alignment of this class-conditional model $p_{\theta}(\mathbf{x} \mid y)$ with that of the ground truth data distribution $p_{\text{data}}(\mathbf{x} \mid y)$. We test this hypothesis by studying the effect of density alignment on gradient explanations. To achieve this density alignment, we use an algorithm called score-matching, and propose novel approximations to this algorithm to enable training large-scale models.

Our experiments show that improving the alignment of the implicit density model with the data distribution enhances gradient structure and explanatory power while reducing this alignment has the opposite effect. This also leads us to conjecture that unintended density alignment in standard neural network training may explain the highly structured nature of input-gradients observed in practice. Overall, our finding that input-gradients capture information regarding an implicit generative model implies that we need to re-think their use for interpreting discriminative models.

Neural Attention Distillation: Erasing Backdoor Triggers from Deep Neural Networks

- Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, Xingjun Ma
- abstract@open-review(Poster): Deep neural networks (DNNs) are known vulnerable to backdoor attacks, a training time attack that injects a trigger pattern into a small proportion of training data so as to control the model's prediction at the test time. Backdoor attacks are notably dangerous since they do not affect the model's performance on clean examples, yet can fool the model to make the incorrect prediction whenever the trigger pattern appears during testing. In this paper, we propose a novel defense framework Neural Attention Distillation (NAD) to erase backdoor triggers from backdoored DNNs. NAD utilizes a teacher network to guide the finetuning of the backdoored student network on a small clean subset of data such that the intermediate-layer attention of the student network aligns with that of the teacher network. The teacher network can be obtained by an independent finetuning process on the same clean subset. We empirically show, against 6 state-of-the-art backdoor attacks, NAD can effectively erase the backdoor triggers using only 5\% clean training data without causing obvious performance degradation on clean examples. Our code is available at https://github.com/bboylyg/NAD.

Data-Efficient Reinforcement Learning with Self-Predictive Representations

- Max Schwarzer, Ankesh Anand, Rishab Goel, R Devon Hjelm, Aaron Courville, Philip Bachman
- abstract@open-review(Spotlight): While deep reinforcement learning excels at solving tasks where large amounts of data can be collected through virtually unlimited interaction with the environment, learning from limited interaction remains a key challenge. We posit that an agent can learn more efficiently if we augment reward maximization with self-supervised objectives based on structure in its visual input and sequential interaction with the environment. Our method, Self-Predictive Representations (SPR), trains an agent to predict its own latent state representations multiple steps into the future. We compute target representations for future states using an encoder which is an exponential moving average of the agent's parameters and we make predictions using a learned transition model. On its own, this future prediction objective outperforms prior methods for sample-efficient deep RL from pixels. We further improve performance by adding data augmentation to the future prediction loss, which forces the agent's representations to be consistent across multiple views of an observation. Our full self-supervised objective, which combines future prediction and data augmentation, achieves a median human-normalized score of 0.415 on Atari in a setting limited to 100k steps of environment interaction, which represents a 55% relative improvement over the previous state-of-the-art. Notably, even in this limited data regime, SPR exceeds expert human scores on 7 out of 26 games. We've made the code associated with this work available at https://github.com/mila-iqia/spr.

HyperGrid Transformers: Towards A Single Model for Multiple Tasks

- Yi Tay, Zhe Zhao, Dara Bahri, Donald Metzler, Da-Cheng Juan
- abstract@open-review(Poster): Achieving state-of-the-art performance on natural language understanding tasks typically relies on fine-tuning a fresh model for every task. Consequently, this approach leads to a higher overall parameter cost, along with higher technical maintenance for serving multiple models. Learning a single multi-task model that is able to do well for all the tasks has been a challenging and yet attractive proposition. In this paper, we propose HyperGrid Transformers, a new Transformer architecture that leverages task-conditioned hyper networks for controlling its feed-forward layers. Specifically, we propose a decomposable hypernetwork that learns grid-wise projections that help to specialize regions in weight matrices for different tasks. In order to construct the proposed hypernetwork, our method learns the interactions and composition between a global (task-agnostic) state and a local task-specific state. We conduct an extensive set of experiments on GLUE/SuperGLUE. On the SuperGLUE test set, we match the performance of the

state-of-the-art while being 16 times more parameter efficient. Our method helps bridge the gap between fine-tuning and multi-task learning approaches.

Long Range Arena : A Benchmark for Efficient Transformers

- Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, Donald Metzler
- abstract@open-review(Poster): Transformers do not scale very well to long sequence lengths largely because of quadratic self-attention complexity. In the recent months, a wide spectrum of efficient, fast Transformers have been proposed to tackle this problem, more often than not claiming superior or comparable model quality to vanilla Transformer models. To this date, there is no well-established consensus on how to evaluate this class of models. Moreover, inconsistent benchmarking on a wide spectrum of tasks and datasets makes it difficult to assess relative model quality amongst many models. This paper proposes a systematic and unified benchmark, Long Range Arena, specifically focused on evaluating model quality under long-context scenarios. Our benchmark is a suite of tasks consisting of sequences ranging from $1K$ to $16K$ tokens, encompassing a wide range of data types and modalities such as text, natural, synthetic images, and mathematical expressions requiring similarity, structural, and visual-spatial reasoning. We systematically evaluate ten well-established long-range Transformer models (Reformers, Linformers, Linear Transformers, Sinkhorn Transformers, Performers, Synthesizers, Sparse Transformers, and Longformers) on our newly proposed benchmark suite. Long Range Arena paves the way towards better understanding this class of efficient Transformer models, facilitates more research in this direction, and presents new challenging tasks to tackle.

Acting in Delayed Environments with Non-Stationary Markov Policies

- Esther Derman, Gal Dalal, Shie Mannor
- abstract@open-review(Poster): The standard Markov Decision Process (MDP) formulation hinges on the assumption that an action is executed immediately after it was chosen. However, assuming it is often unrealistic and can lead to catastrophic failures in applications such as robotic manipulation, cloud computing, and finance. We introduce a framework for learning and planning in MDPs where the decision-maker commits actions that are executed with a delay of m steps. The brute-force state augmentation baseline where the state is concatenated to the last m committed actions suffers from an exponential complexity in m, as we show for policy iteration. We then prove that with execution delay, deterministic Markov policies in the original state-space are sufficient for attaining maximal reward, but need to be non-stationary. As for stationary Markov policies, we show they are sub-optimal in general. Consequently, we devise a non-stationary Q-learning style model-based algorithm that solves delayed execution tasks without resorting to state-augmentation. Experiments on tabular, physical, and Atari domains reveal that it converges quickly to high performance even for substantial delays, while standard approaches that either ignore the delay or rely on state-augmentation struggle or fail due to divergence. The code is available at \url{https://github.com/galdl/rl_delay_basic.git}.

Neurally Augmented ALISTA

- Freya Behrens, Jonathan Sauder, Peter Jung
- abstract@open-review(Poster): It is well-established that many iterative sparse reconstruction algorithms can be unrolled to yield a learnable neural network for improved empirical performance. A prime example is learned ISTA (LISTA) where weights, step sizes and thresholds are learned from training data. Recently, Analytic LISTA (ALISTA) has been introduced, combining the strong empirical performance of a fully learned approach like LISTA, while retaining theoretical guarantees of classical compressed sensing algorithms and significantly reducing the number of parameters to learn. However, these parameters are trained to work in expectation, often leading to suboptimal reconstruction of individual targets. In this work we therefore introduce Neurally Augmented ALISTA, in which an LSTM network is used to compute step sizes and thresholds individually for each target vector during reconstruction. This adaptive approach is theoretically motivated by revisiting the recovery guarantees of ALISTA. We show that our approach further improves empirical performance in sparse reconstruction, in particular outperforming existing algorithms by an increasing margin as the compression ratio becomes more challenging.

Relating by Contrasting: A Data-efficient Framework for Multimodal Generative Models

- Yuge Shi, Brooks Paige, Philip Torr, Siddharth N
- abstract@open-review(Poster): Multimodal learning for generative models often refers to the learning of abstract concepts from the commonality of information in multiple modalities, such as vision and language. While it has proven effective for learning generalisable representations, the training of such models often requires a large amount of related multimodal data that shares commonality, which can be expensive to come by. To mitigate this, we develop a novel contrastive framework for generative model learning, allowing us to train the model not just by the commonality between modalities, but by the distinction between "related" and "unrelated" multimodal data. We show in experiments that our method enables data-efficient multimodal learning on challenging datasets for various multimodal VAE models. We also show that under our proposed framework, the generative model can accurately identify related samples from unrelated ones, making it possible to make use of the plentiful unlabeled, unpaired multimodal data.

Categorical Normalizing Flows via Continuous Transformations

- Phillip Lippe, Efstratios Gavves
- abstract@open-review(Poster): Despite their popularity, to date, the application of normalizing flows on categorical data stays limited. The current practice of using dequantization to map discrete data to a continuous space is inapplicable as categorical data has no intrinsic order. Instead, categorical data have complex and latent relations that must be inferred, like the synonymy between words. In this paper, we investigate Categorical Normalizing Flows, that is normalizing flows for categorical data. By casting the encoding of categorical data in continuous space as a variational inference problem, we jointly optimize the continuous representation and the model likelihood. Using a factorized decoder, we introduce an inductive bias to model any interactions in the normalizing flow. As a consequence, we do not only simplify the optimization compared to having a joint decoder, but also make it possible to scale up to a large number of categories that is currently impossible with discrete normalizing flows. Based on Categorical Normalizing Flows, we propose GraphCNF a permutation-invariant generative model on graphs. GraphCNF implements a three step approach modeling the nodes, edges, and adjacency matrix stepwise to increase efficiency. On molecule generation, GraphCNF outperforms both one-shot and autoregressive flow-based state-of-the-art.

Prototypical Representation Learning for Relation Extraction

- Ning Ding, Xiaobin Wang, Yao Fu, Guangwei Xu, Rui Wang, Pengjun Xie, Ying Shen, Fei Huang, Hai-Tao Zheng, Rui Zhang
- abstract@open-review(Poster): Recognizing relations between entities is a pivotal task of relational learning.
  Learning relation representations from distantly-labeled datasets is difficult because of the abundant label noise and complicated expressions in human language.
  This paper aims to learn predictive, interpretable, and robust relation representations from distantly-labeled data that are effective in different settings, including supervised, distantly supervised, and few-shot learning. Instead of solely relying on the supervision from noisy labels, we propose to learn prototypes for each relation from contextual information to best explore the intrinsic semantics of relations. Prototypes are representations in the feature space abstracting the essential semantics of relations between entities in sentences. We learn prototypes based on objectives with clear geometric interpretation, where the prototypes are unit vectors uniformly dispersed in a unit ball, and statement embeddings are centered at the end of their corresponding prototype vectors on the surface of the ball. This approach allows us to learn meaningful, interpretable prototypes for the final

classification. Results on several relation learning tasks show that our model significantly outperforms the previous state-of-the-art models. We further demonstrate the robustness of the encoder and the interpretability of prototypes with extensive experiments.

Generalized Energy Based Models

- Michael Arbel, Liang Zhou, Arthur Gretton
- abstract@open-review(Poster): We introduce the Generalized Energy Based Model (GEBM) for generative modelling. These models combine two trained components: a base distribution (generally an implicit model), which can learn the support of data with low intrinsic dimension in a high dimensional space; and an energy function, to refine the probability mass on the learned support. Both the energy function and base jointly constitute the final model, unlike GANs, which retain only the base distribution (the "generator").
  GEBMs are trained by alternating between learning the energy and the base. We show that both training stages are well-defined: the energy is learned by maximising a generalized likelihood, and the resulting energy-based loss provides informative gradients for learning the base. Samples from the posterior on the latent space of the trained model can be obtained via MCMC, thus finding regions in this space that produce better quality samples. Empirically, the GEBM samples on image-generation tasks are of much better quality than those from the learned generator alone, indicating that all else being equal, the GEBM will outperform a GAN of the same complexity. When using normalizing flows as base measures, GEBMs succeed on density modelling tasks returning comparable performance to direct maximum likelihood of the same networks.

Predicting Classification Accuracy When Adding New Unobserved Classes

- Yuli Slavutsky, Yuval Benjamini
- abstract@open-review(Poster): Multiclass classifiers are often designed and evaluated only on a sample from the classes on which they will eventually be applied. Hence, their final accuracy remains unknown. In this work we study how a classifier’s performance over the initial class sample can be used to extrapolate its expected accuracy on a larger, unobserved set of classes. For this, we define a measure of separation between correct and incorrect classes that is independent of the number of classes: the "reversed ROC" (rROC), which is obtained by replacing the roles of classes and data-points in the common ROC. We show that the classification accuracy is a function of the rROC in multiclass classifiers, for which the learned representation of data from the initial class sample remains unchanged when new classes are added. Using these results we formulate a robust neural-network-based algorithm, "CleaneX", which learns to estimate the accuracy of such classifiers on arbitrarily large sets of classes. Unlike previous methods, our method uses both the observed accuracies of the classifier and densities of classification scores, and therefore achieves remarkably better predictions than current state-of-the-art methods on both simulations and real datasets of object detection, face recognition, and brain decoding.

In Search of Lost Domain Generalization

- Ishaan Gulrajani, David Lopez-Paz
- abstract@open-review(Poster): The goal of domain generalization algorithms is to predict well on distributions different from those seen during training. While a myriad of domain generalization algorithms exist, inconsistencies in experimental conditions---datasets, network architectures, and model selection criteria---render fair comparisons difficult. The goal of this paper is to understand how useful domain generalization algorithms are in realistic settings. As a first step, we realize that model selection is non-trivial for domain generalization tasks, and we argue that algorithms without a model selection criterion remain incomplete. Next we implement DomainBed, a testbed for domain generalization including seven benchmarks, fourteen algorithms, and three model selection criteria. When conducting extensive experiments using DomainBed we find that when carefully implemented and tuned, ERM outperforms the state-of-the-art in terms of average performance. Furthermore, no algorithm included in DomainBed outperforms ERM by more than one point when evaluated under the same experimental conditions. We hope that the release of DomainBed, alongside contributions from fellow researchers, will streamline reproducible and rigorous advances in domain generalization.

Estimating informativeness of samples with Smooth Unique Information

- Hrayr Harutyunyan, Alessandro Achille, Giovanni Paolini, Orchid Majumder, Avinash Ravichandran, Rahul Bhotika, Stefano Soatto
- abstract@open-review(Poster): We define a notion of information that an individual sample provides to the training of a neural network, and we specialize it to measure both how much a sample informs the final weights and how much it informs the function computed by the weights. Though related, we show that these quantities have a qualitatively different behavior. We give efficient approximations of these quantities using a linearized network and demonstrate empirically that the approximation is accurate for real-world architectures, such as pre-trained ResNets. We apply these measures to several problems, such as dataset summarization, analysis of under-sampled classes, comparison of informativeness of different data sources, and detection of adversarial and corrupted examples. Our work generalizes existing frameworks, but enjoys better computational properties for heavily over-parametrized models, which makes it possible to apply it to real-world networks.

Identifying Physical Law of Hamiltonian Systems via Meta-Learning

- Seungjun Lee, Haesang Yang, Woojae Seong
- abstract@open-review(Poster): Hamiltonian mechanics is an effective tool to represent many physical processes with concise yet well-generalized mathematical expressions. A well-modeled Hamiltonian makes it easy for researchers to analyze and forecast many related phenomena that are governed by the same physical law. However, in general, identifying a functional or shared expression of the Hamiltonian is very difficult. It requires carefully designed experiments and the researcher's insight that comes from years of experience. We propose that meta-learning algorithms can be potentially powerful data-driven tools for identifying the physical law governing Hamiltonian systems without any mathematical assumptions on the representation, but with observations from a set of systems governed by the same physical law. We show that a well meta-trained learner can identify the shared representation of the Hamiltonian by evaluating our method on several types of physical systems with various experimental settings.

Adapting to Reward Progressivity via Spectral Reinforcement Learning

- Michael Dann, John Thangarajah
- abstract@open-review(Poster): In this paper we consider reinforcement learning tasks with progressive rewards; that is, tasks where the rewards tend to increase in magnitude over time. We hypothesise that this property may be problematic for value-based deep reinforcement learning agents, particularly if the agent must first succeed in relatively unrewarding regions of the task in order to reach more rewarding regions. To address this issue, we propose Spectral DQN, which decomposes the reward into frequencies such that the high frequencies only activate when large rewards are found. This allows the training loss to be balanced so that it gives more even weighting across small and large reward regions. In two domains with extreme reward progressivity, where standard value-based methods struggle significantly, Spectral DQN is able to make much farther progress. Moreover, when evaluated on a set of six standard Atari games that do not overtly favour the approach, Spectral DQN remains more than competitive: While it underperforms one of the benchmarks in a single game, it comfortably surpasses the benchmarks in three games. These results demonstrate that the approach is not overfit to its target problem, and suggest that Spectral DQN may have advantages beyond addressing reward progressivity.

Understanding the effects of data parallelism and sparsity on neural network training

- Namhoon Lee, Thalaiyasingam Ajanthan, Philip Torr, Martin Jaggi

- abstract@open-review(Poster): We study two factors in neural network training: data parallelism and sparsity; here, data parallelism means processing training data in parallel using distributed systems (or equivalently increasing batch size), so that training can be accelerated; for sparsity, we refer to pruning parameters in a neural network model, so as to reduce computational and memory cost. Despite their promising benefits, however, understanding of their effects on neural network training remains elusive. In this work, we first measure these effects rigorously by conducting extensive experiments while tuning all metaparameters involved in the optimization. As a result, we find across various workloads of data set, network model, and optimization algorithm that there exists a general scaling trend between batch size and number of training steps to convergence for the effect of data parallelism, and further, difficulty of training under sparsity. Then, we develop a theoretical analysis based on the convergence properties of stochastic gradient methods and smoothness of the optimization landscape, which illustrates the observed phenomena precisely and generally, establishing a better account of the effects of data parallelism and sparsity on neural network training.

Primal Wasserstein Imitation Learning

- Robert Dadashi, Leonard Hussenot, Matthieu Geist, Olivier Pietquin
- abstract@open-review(Poster): Imitation Learning (IL) methods seek to match the behavior of an agent with that of an expert. In the present work, we propose a new IL method based on a conceptually simple algorithm: Primal Wasserstein Imitation Learning (PWIL), which ties to the primal form of the Wasserstein distance between the expert and the agent state-action distributions. We present a reward function which is derived offline, as opposed to recent adversarial IL algorithms that learn a reward function through interactions with the environment, and which requires little fine-tuning. We show that we can recover expert behavior on a variety of continuous control tasks of the MuJoCo domain in a sample efficient manner in terms of agent interactions and of expert interactions with the environment. Finally, we show that the behavior of the agent we train matches the behavior of the expert with the Wasserstein distance, rather than the commonly used proxy of performance.

Prediction and generalisation over directed actions by grid cells

- Changmin Yu, Timothy Behrens, Neil Burgess
- abstract@open-review(Poster): Knowing how the effects of directed actions generalise to new situations (e.g. moving North, South, East and West, or turning left, right, etc.) is key to rapid generalisation across new situations. Markovian tasks can be characterised by a state space and a transition matrix and recent work has proposed that neural grid codes provide an efficient representation of the state space, as eigenvectors of a transition matrix reflecting diffusion across states, that allows efficient prediction of future state distributions. Here we extend the eigenbasis prediction model, utilising tools from Fourier analysis, to prediction over arbitrary translation-invariant directed transition structures (i.e. displacement and diffusion), showing that a single set of eigenvectors can support predictions over arbitrary directed actions via action-specific eigenvalues. We show how to define a "sense of direction" to combine actions to reach a target state (ignoring task-specific deviations from translation-invariance), and demonstrate that adding the Fourier representations to a deep Q network aids policy learning in continuous control tasks. We show the equivalence between the generalised prediction framework and traditional models of grid cell firing driven by self-motion to perform path integration, either using oscillatory interference (via Fourier components as velocity-controlled oscillators) or continuous attractor networks (via analysis of the update dynamics). We thus provide a unifying framework for the role of the grid system in predictive planning, sense of direction and path integration: supporting generalisable inference over directed actions across different tasks.

Drop-Bottleneck: Learning Discrete Compressed Representation for Noise-Robust Exploration

- Jaekyeom Kim, Minjung Kim, Dongyeon Woo, Gunhee Kim
- abstract@open-review(Poster): We propose a novel information bottleneck (IB) method named Drop-Bottleneck, which discretely drops features that are irrelevant to the target variable. Drop-Bottleneck not only enjoys a simple and tractable compression objective but also additionally provides a deterministic compressed representation of the input variable, which is useful for inference tasks that require consistent representation. Moreover, it can jointly learn a feature extractor and select features considering each feature dimension's relevance to the target task, which is unattainable by most neural network-based IB methods. We propose an exploration method based on Drop-Bottleneck for reinforcement learning tasks. In a multitude of noisy and reward sparse maze navigation tasks in VizDoom (Kempka et al., 2016) and DMLab (Beattie et al., 2016), our exploration method achieves state-of-the-art performance. As a new IB framework, we demonstrate that Drop-Bottleneck outperforms Variational Information Bottleneck (VIB) (Alemi et al., 2017) in multiple aspects including adversarial robustness and dimensionality reduction.

BOIL: Towards Representation Change for Few-shot Learning

- Jaehoon Oh, Hyungjun Yoo, ChangHwan Kim, Se-Young Yun
- abstract@open-review(Poster): Model Agnostic Meta-Learning (MAML) is one of the most representative of gradient-based meta-learning algorithms. MAML learns new tasks with a few data samples using inner updates from a meta-initialization point and learns the meta-initialization parameters with outer updates. It has recently been hypothesized that representation reuse, which makes little change in efficient representations, is the dominant factor in the performance of the meta-initialized model through MAML in contrast to representation change, which causes a significant change in representations. In this study, we investigate the necessity of representation change for the ultimate goal of few-shot learning, which is solving domain-agnostic tasks. To this aim, we propose a novel meta-learning algorithm, called BOIL (Body Only update in Inner Loop), which updates only the body (extractor) of the model and freezes the head (classifier) during inner loop updates. BOIL leverages representation change rather than representation reuse. A frozen head cannot achieve better results than even a random guessing classifier at the initial point of new tasks, and feature vectors (representations) have to move quickly to their corresponding frozen head vectors. We visualize this property using cosine similarity, CKA, and empirical results without the head. Although the inner loop updates purely hinge on representation change, BOIL empirically shows significant performance improvement over MAML, particularly on cross-domain tasks. The results imply that representation change in gradient-based meta-learning approaches is a critical component.

MultiModalQA: complex question answering over text, tables and images

- Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, Jonathan Berant
- abstract@open-review(Poster): When answering complex questions, people can seamlessly combine information from visual, textual and tabular sources. While interest in models that reason over multiple pieces of evidence has surged in recent years, there has been relatively little work on question answering models that reason across multiple modalities. In this paper, we present MultiModalQA (MMQA): a challenging question answering dataset that requires joint reasoning over text, tables and images. We create MMQA using a new framework for generating complex multi-modal questions at scale, harvesting tables from Wikipedia, and attaching images and text paragraphs using entities that appear in each table. We then define a formal language that allows us to take questions that can be answered from a single modality, and combine them to generate cross-modal questions. Last, crowdsourcing workers take these automatically generated questions and rephrase them into more fluent language. We create 29,918 questions through this procedure, and empirically demonstrate the necessity of a multi-modal multi-hop approach to solve our task: our multi-hop model, ImplicitDecomp, achieves an average F1 of 51.7 over cross-modal questions, substantially outperforming a strong baseline that achieves 38.2 F1, but still lags significantly behind human performance, which is at 90.1 F1.

Neural gradients are near-lognormal: improved quantized and sparse training

- Brian Chmiel, Liad Ben-Uri, Moran Shkolnik, Elad Hoffer, Ron Banner, Daniel Soudry

- abstract@open-review(Poster): While training can mostly be accelerated by reducing the time needed to propagate neural gradients (loss gradients with respect to the intermediate neural layer outputs) back throughout the model, most previous works focus on the quantization/pruning of weights and activations. These methods are often not applicable to neural gradients, which have very different statistical properties. Distinguished from weights and activations, we find that the distribution of neural gradients is approximately lognormal. Considering this, we suggest two closed-form analytical methods to reduce the computational and memory burdens of neural gradients. The first method optimizes the floating-point format and scale of the gradients. The second method accurately sets sparsity thresholds for gradient pruning. Each method achieves state-of-the-art results on ImageNet. To the best of our knowledge, this paper is the first to (1) quantize the gradients to 6-bit floating-point formats, or (2) achieve up to 85% gradient sparsity --- in each case without accuracy degradation. Reference implementation accompanies the paper in the supplementary material.

Group Equivariant Conditional Neural Processes

- Makoto Kawano, Wataru Kumagai, Akiyoshi Sannai, Yusuke Iwasawa, Yutaka Matsuo
- abstract@open-review(Poster): We present the group equivariant conditional neural process (EquivCNP), a meta-learning method with permutation invariance in a data set as in conventional conditional neural processes (CNPs), and it also has transformation equivariance in data space. Incorporating group equivariance, such as rotation and scaling equivariance, provides a way to consider the symmetry of real-world data. We give a decomposition theorem for permutation-invariant and group-equivariant maps, which leads us to construct EquivCNPs with an infinite-dimensional latent space to handle group symmetries. In this paper, we build architecture using Lie group convolutional layers for practical implementation. We show that EquivCNP with translation equivariance achieves comparable performance to conventional CNPs in a 1D regression task. Moreover, we demonstrate that incorporating an appropriate Lie group equivariance, EquivCNP is capable of zero-shot generalization for an image-completion task by selecting an appropriate Lie group equivariance.

Deployment-Efficient Reinforcement Learning via Model-Based Offline Optimization

- Tatsuya Matsushima, Hiroki Furuta, Yutaka Matsuo, Ofir Nachum, Shixiang Gu
- abstract@open-review(Poster): Most reinforcement learning (RL) algorithms assume online access to the environment, in which one may readily interleave updates to the policy with experience collection using that policy. However, in many real-world applications such as health, education, dialogue agents, and robotics, the cost or potential risk of deploying a new data-collection policy is high, to the point that it can become prohibitive to update the data-collection policy more than a few times during learning. With this view, we propose a novel concept of deployment efficiency, measuring the number of distinct data-collection policies that are used during policy learning. We observe that naïvely applying existing model-free offline RL algorithms recursively does not lead to a practical deployment-efficient and sample-efficient algorithm. We propose a novel model-based algorithm, Behavior-Regularized Model-ENsemble (BREMEN), that not only performs better than or comparably as the state-of-the-art dynamic-programming-based and concurrently-proposed model-based offline approaches on existing benchmarks, but can also effectively optimize a policy offline using 10-20 times fewer data than prior works. Furthermore, the recursive application of BREMEN achieves impressive deployment efficiency while maintaining the same or better sample efficiency, learning successful policies from scratch on simulated robotic environments with only 5-10 deployments, compared to typical values of hundreds to millions in standard RL baselines.

Efficient Conformal Prediction via Cascaded Inference with Expanded Admission

- Adam Fisch, Tal Schuster, Tommi S. Jaakkola, Regina Barzilay
- abstract@open-review(Poster): In this paper, we present a novel approach for conformal prediction (CP), in which we aim to identify a set of promising prediction candidates---in place of a single prediction. This set is guaranteed to contain a correct answer with high probability, and is well-suited for many open-ended classification tasks. In the standard CP paradigm, the predicted set can often be unusably large and also costly to obtain. This is particularly pervasive in settings where the correct answer is not unique, and the number of total possible answers is high. We first expand the CP correctness criterion to allow for additional, inferred "admissible" answers, which can substantially reduce the size of the predicted set while still providing valid performance guarantees. Second, we amortize costs by conformalizing prediction cascades, in which we aggressively prune implausible labels early on by using progressively stronger classifiers---again, while still providing valid performance guarantees. We demonstrate the empirical effectiveness of our approach for multiple applications in natural language processing and computational chemistry for drug discovery.

Reducing the Computational Cost of Deep Generative Models with Binary Neural Networks

- Thomas Bird, Friso Kingma, David Barber
- abstract@open-review(Poster): Deep generative models provide a powerful set of tools to understand real-world data. But as these models improve, they increase in size and complexity, so their computational cost in memory and execution time grows. Using binary weights in neural networks is one method which has shown promise in reducing this cost. However, whether binary neural networks can be used in generative models is an open problem. In this work we show, for the first time, that we can successfully train generative models which utilize binary neural networks. This reduces the computational cost of the models massively. We develop a new class of binary weight normalization, and provide insights for architecture designs of these binarized generative models. We demonstrate that two state-of-the-art deep generative models, the ResNet VAE and Flow++ models, can be binarized effectively using these techniques. We train binary models that achieve loss values close to those of the regular models but are 90%-94% smaller in size, and also allow significant speed-ups in execution time.

Large-width functional asymptotics for deep Gaussian neural networks

- Daniele Bracale, Stefano Favaro, Sandra Fortini, Stefano Peluchetti
- abstract@open-review(Poster): In this paper, we consider fully connected feed-forward deep neural networks where weights and biases are independent and identically distributed according to Gaussian distributions. Extending previous results (Matthews et al., 2018a;b;Yang, 2019) we adopt a function-space perspective, i.e. we look at neural networks as infinite-dimensional random elements on the input space $\mathbb{R}^I$. Under suitable assumptions on the activation function we show that: i) a network defines a continuous Gaussian process on the input space $\mathbb{R}^I$; ii) a network with re-scaled weights converges weakly to a continuous Gaussian process in the large-width limit; iii) the limiting Gaussian process has almost surely locally γ-Hölder continuous paths, for $0 < \gamma <1$. Our results contribute to recent theoretical studies on the interplay between infinitely wide deep neural networks and Gaussian processes by establishing weak convergence in function-space with respect to a stronger metric.

MetaNorm: Learning to Normalize Few-Shot Batches Across Domains

- Yingjun Du, Xiantong Zhen, Ling Shao, Cees G. M. Snoek
- abstract@open-review(Poster): Batch normalization plays a crucial role when training deep neural networks. However, batch statistics become unstable with small batch sizes and are unreliable in the presence of distribution shifts. We propose MetaNorm, a simple yet effective meta-learning normalization. It tackles the aforementioned issues in a unified way by leveraging the meta-learning setting and learns to infer adaptive statistics for batch normalization. MetaNorm is generic, flexible and model-agnostic, making it a simple plug-and-play module that is seamlessly embedded into existing meta-learning approaches. It can be efficiently implemented by lightweight hypernetworks with low computational cost. We verify its effectiveness by extensive evaluation on representative tasks suffering from the small batch and domain shift problems: few-shot learning and domain generalization. We further introduce an even more challenging setting: few-shot domain generalization. Results demonstrate that MetaNorm consistently achieves better, or at least competitive, accuracy compared to existing batch normalization methods.

Representation Balancing Offline Model-based Reinforcement Learning

- Byung-Jun Lee, Jongmin Lee, Kee-Eung Kim
- abstract@open-review(Poster): One of the main challenges in offline and off-policy reinforcement learning is to cope with the distribution shift that arises from the mismatch between the target policy and the data collection policy. In this paper, we focus on a model-based approach, particularly on learning the representation for a robust model of the environment under the distribution shift, which has been first studied by Representation Balancing MDP (RepBM). Although this prior work has shown promising results, there are a number of shortcomings that still hinder its applicability to practical tasks. In particular, we address the curse of horizon exhibited by RepBM, rejecting most of the pre-collected data in long-term tasks. We present a new objective for model learning motivated by recent advances in the estimation of stationary distribution corrections. This effectively overcomes the aforementioned limitation of RepBM, as well as naturally extending to continuous action spaces and stochastic policies. We also present an offline model-based policy optimization using this new objective, yielding the state-of-the-art performance in a representative set of benchmark offline RL tasks.

Local Convergence Analysis of Gradient Descent Ascent with Finite Timescale Separation

- Tanner Fiez, Lillian J Ratliff
- abstract@open-review(Poster): We study the role that a finite timescale separation parameter τ has on gradient descent-ascent in non-convex, non-concave zero-sum games where the learning rate of player 1 is denoted by γ_1 and the learning rate of player 2 is defined to be $\gamma_2=\tau\gamma_1$. We provide a non-asymptotic construction of the finite timescale separation parameter $\tau^{\ast}$ such that gradient descent-ascent locally converges to $x^{\ast}$ for all $\tau \in (\tau^{\ast}, \infty)$ if and only if it is a strict local minmax equilibrium. Moreover, we provide explicit local convergence rates given the finite timescale separation. The convergence results we present are complemented by a non-convergence result: given a critical point $x^{\ast}$ that is not a strict local minmax equilibrium, we present a non-asymptotic construction of a finite timescale separation τ_{0} such that gradient descent-ascent with timescale separation $\tau\in (\tau_0, \infty)$ does not converge to $x^{\ast}$. Finally, we extend the results to gradient penalty regularization methods for generative adversarial networks and empirically demonstrate on CIFAR-10 and CelebA the significant impact timescale separation has on training performance.

Meta-GMVAE: Mixture of Gaussian VAE for Unsupervised Meta-Learning

- Dong Bok Lee, Dongchan Min, Seanie Lee, Sung Ju Hwang
- abstract@open-review(Spotlight): Unsupervised learning aims to learn meaningful representations from unlabeled data which can captures its intrinsic structure, that can be transferred to downstream tasks. Meta-learning, whose objective is to learn to generalize across tasks such that the learned model can rapidly adapt to a novel task, shares the spirit of unsupervised learning in that the both seek to learn more effective and efficient learning procedure than learning from scratch. The fundamental difference of the two is that the most meta-learning approaches are supervised, assuming full access to the labels. However, acquiring labeled dataset for meta-training not only is costly as it requires human efforts in labeling but also limits its applications to pre-defined task distributions. In this paper, we propose a principled unsupervised meta-learning model, namely Meta-GMVAE, based on Variational Autoencoder (VAE) and set-level variational inference. Moreover, we introduce a mixture of Gaussian (GMM) prior, assuming that each modality represents each class-concept in a randomly sampled episode, which we optimize with Expectation-Maximization (EM). Then, the learned model can be used for downstream few-shot classification tasks, where we obtain task-specific parameters by performing semi-supervised EM on the latent representations of the support and query set, and predict labels of the query set by computing aggregated posteriors. We validate our model on Omniglot and Mini-ImageNet datasets by evaluating its performance on downstream few-shot classification tasks. The results show that our model obtain impressive performance gains over existing unsupervised meta-learning baselines, even outperforming supervised MAML on a certain setting.

The Importance of Pessimism in Fixed-Dataset Policy Optimization

- Jacob Buckman, Carles Gelada, Marc G Bellemare
- abstract@open-review(Poster): We study worst-case guarantees on the expected return of fixed-dataset policy optimization algorithms. Our core contribution is a unified conceptual and mathematical framework for the study of algorithms in this regime. This analysis reveals that for naive approaches, the possibility of erroneous value overestimation leads to a difficult-to-satisfy requirement: in order to guarantee that we select a policy which is near-optimal, we may need the dataset to be informative of the value of every policy. To avoid this, algorithms can follow the pessimism principle, which states that we should choose the policy which acts optimally in the worst possible world. We show why pessimistic algorithms can achieve good performance even when the dataset is not informative of every policy, and derive families of algorithms which follow this principle. These theoretical findings are validated by experiments on a tabular gridworld, and deep learning experiments on four MinAtar environments.

Uncertainty Estimation and Calibration with Finite-State Probabilistic RNNs

- Cheng Wang, Carolin Lawrence, Mathias Niepert
- abstract@open-review(Poster): Uncertainty quantification is crucial for building reliable and trustable machine learning systems. We propose to estimate uncertainty in recurrent neural networks (RNNs) via stochastic discrete state transitions over recurrent timesteps. The uncertainty of the model can be quantified by running a prediction several times, each time sampling from the recurrent state transition distribution, leading to potentially different results if the model is uncertain. Alongside uncertainty quantification, our proposed method offers several advantages in different settings. The proposed method can (1) learn deterministic and probabilistic automata from data, (2) learn well-calibrated models on real-world classification tasks, (3) improve the performance of out-of-distribution detection, and (4) control the exploration-exploitation trade-off in reinforcement learning. An implementation is available.

Empirical or Invariant Risk Minimization? A Sample Complexity Perspective

- Kartik Ahuja, Jun Wang, Amit Dhurandhar, Karthikeyan Shanmugam, Kush R. Varshney
- abstract@open-review(Poster): Recently, invariant risk minimization (IRM) was proposed as a promising solution to address out-of-distribution (OOD) generalization. However, it is unclear when IRM should be preferred over the widely-employed empirical risk minimization (ERM) framework. In this work, we analyze both these frameworks from the perspective of sample complexity, thus taking a firm step towards answering this important question. We find that depending on the type of data generation mechanism, the two approaches might have very different finite sample and asymptotic behavior. For example, in the covariate shift setting we see that the two approaches not only arrive at the same asymptotic solution, but also have similar finite sample behavior with no clear winner. For other distribution shifts such as those involving confounders or anti-causal variables, however, the two approaches arrive at different asymptotic solutions where IRM is guaranteed to be close to the desired OOD solutions in the finite sample regime, while ERM is biased even asymptotically. We further investigate how different factors --- the number of environments, complexity of the model, and IRM penalty weight --- impact the sample complexity of IRM in relation to its distance from the OOD solutions.

Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval

- Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, Arnold Overwijk
- abstract@open-review(Poster): Conducting text retrieval in a learned dense representation space has many intriguing advantages. Yet dense retrieval (DR) often underperforms word-based sparse retrieval. In this paper, we first theoretically show the bottleneck of dense retrieval is the domination of

uninformative negatives sampled in mini-batch training, which yield diminishing gradient norms, large gradient variances, and slow convergence. We then propose Approximate nearest neighbor Negative Contrastive Learning (ANCE), which selects hard training negatives globally from the entire corpus. Our experiments demonstrate the effectiveness of ANCE on web search, question answering, and in a commercial search engine, showing ANCE dot-product retrieval nearly matches the accuracy of BERT-based cascade IR pipeline. We also empirically validate our theory that negative sampling with ANCE better approximates the oracle importance sampling procedure and improves learning convergence.

Implicit Convex Regularizers of CNN Architectures: Convex Optimization of Two- and Three-Layer Networks in Polynomial Time

- Tolga Ergen, Mert Pilanci
- abstract@open-review(Spotlight): We study training of Convolutional Neural Networks (CNNs) with ReLU activations and introduce exact convex optimization formulations with a polynomial complexity with respect to the number of data samples, the number of neurons, and data dimension. More specifically, we develop a convex analytic framework utilizing semi-infinite duality to obtain equivalent convex optimization problems for several two- and three-layer CNN architectures. We first prove that two-layer CNNs can be globally optimized via an ℓ_2 norm regularized convex program. We then show that multi-layer circular CNN training problems with a single ReLU layer are equivalent to an ℓ_1 regularized convex program that encourages sparsity in the spectral domain. We also extend these results to three-layer CNNs with two ReLU layers. Furthermore, we present extensions of our approach to different pooling methods, which elucidates the implicit architectural bias as convex regularizers.

FairBatch: Batch Selection for Model Fairness

- Yuji Roh, Kangwook Lee, Steven Euijong Whang, Changho Suh
- abstract@open-review(Poster): Training a fair machine learning model is essential to prevent demographic disparity. Existing techniques for improving model fairness require broad changes in either data preprocessing or model training, rendering themselves difficult-to-adopt for potentially already complex machine learning systems. We address this problem via the lens of bilevel optimization. While keeping the standard training algorithm as an inner optimizer, we incorporate an outer optimizer so as to equip the inner problem with an additional functionality: Adaptively selecting minibatch sizes for the purpose of improving model fairness. Our batch selection algorithm, which we call FairBatch, implements this optimization and supports prominent fairness measures: equal opportunity, equalized odds, and demographic parity. FairBatch comes with a significant implementation benefit -- it does not require any modification to data preprocessing or model training. For instance, a single-line change of PyTorch code for replacing batch selection part of model training suffices to employ FairBatch. Our experiments conducted both on synthetic and benchmark real data demonstrate that FairBatch can provide such functionalities while achieving comparable (or even greater) performances against the state of the arts. Furthermore, FairBatch can readily improve fairness of any pre-trained model simply via fine-tuning. It is also compatible with existing batch selection techniques intended for different purposes, such as faster convergence, thus gracefully achieving multiple purposes.

An Unsupervised Deep Learning Approach for Real-World Image Denoising

- Dihan Zheng, Sia Huat Tan, Xiaowen Zhang, Zuoqiang Shi, Kaisheng Ma, Chenglong Bao
- abstract@open-review(Poster): Designing an unsupervised image denoising approach in practical applications is a challenging task due to the complicated data acquisition process. In the real-world case, the noise distribution is so complex that the simplified additive white Gaussian (AWGN) assumption rarely holds, which significantly deteriorates the Gaussian denoisers' performance. To address this problem, we apply a deep neural network that maps the noisy image into a latent space in which the AWGN assumption holds, and thus any existing Gaussian denoiser is applicable. More specifically, the proposed neural network consists of the encoder-decoder structure and approximates the likelihood term in the Bayesian framework. Together with a Gaussian denoiser, the neural network can be trained with the input image itself and does not require any pre-training in other datasets. Extensive experiments on real-world noisy image datasets have shown that the combination of neural networks and Gaussian denoisers improves the performance of the original Gaussian denoisers by a large margin. In particular, the neural network+BM3D method significantly outperforms other unsupervised denoising approaches and is competitive with supervised networks such as DnCNN, FFDNet, and CBDNet.

When Optimizing f-Divergence is Robust with Label Noise

- Jiaheng Wei, Yang Liu
- abstract@open-review(Poster): We show when maximizing a properly defined f-divergence measure with respect to a classifier's predictions and the supervised labels is robust with label noise. Leveraging its variational form, we derive a nice decoupling property for a family of f-divergence measures when label noise presents, where the divergence is shown to be a linear combination of the variational difference defined on the clean distribution and a bias term introduced due to the noise. The above derivation helps us analyze the robustness of different f-divergence functions. With established robustness, this family of f-divergence functions arises as useful metrics for the problem of learning with noisy labels, which do not require the specification of the labels' noise rate. When they are possibly not robust, we propose fixes to make them so. In addition to the analytical results, we present thorough experimental evidence. Our code is available at https://github.com/UCSC-REAL/Robust-f-divergence-measures.

Meta-learning Symmetries by Reparameterization

- Allan Zhou, Tom Knowles, Chelsea Finn
- abstract@open-review(Poster): Many successful deep learning architectures are equivariant to certain transformations in order to conserve parameters and improve generalization: most famously, convolution layers are equivariant to shifts of the input. This approach only works when practitioners know the symmetries of the task and can manually construct an architecture with the corresponding equivariances. Our goal is an approach for learning equivariances from data, without needing to design custom task-specific architectures. We present a method for learning and encoding equivariances into networks by learning corresponding parameter sharing patterns from data. Our method can provably represent equivariance-inducing parameter sharing for any finite group of symmetry transformations. Our experiments suggest that it can automatically learn to encode equivariances to common transformations used in image processing tasks.

A Geometric Analysis of Deep Generative Image Models and Its Applications

- Binxu Wang, Carlos R Ponce
- abstract@open-review(Poster): Generative adversarial networks (GANs) have emerged as a powerful unsupervised method to model the statistical patterns of real-world data sets, such as natural images. These networks are trained to map random inputs in their latent space to new samples representative of the learned data. However, the structure of the latent space is hard to intuit due to its high dimensionality and the non-linearity of the generator, which limits the usefulness of the models. Understanding the latent space requires a way to identify input codes for existing real-world images (inversion), and a way to identify directions with known image transformations (interpretability). Here, we use a geometric framework to address both issues simultaneously. We develop an architecture-agnostic method to compute the Riemannian metric of the image manifold created by GANs. The eigen-decomposition of the metric isolates axes that account for different levels of image variability. An empirical analysis of several pretrained GANs shows that image variation around each position is concentrated along surprisingly few major axes (the space is highly anisotropic) and the directions that create this large variation are similar at different positions in the space (the space is homogeneous). We show that many of the top eigenvectors correspond to interpretable transforms in the image space, with a substantial part of eigenspace corresponding to minor transforms which could be compressed out. This geometric understanding unifies key previous results related to GAN interpretability. We show that the use of this metric allows for more efficient

optimization in the latent space (e.g. GAN inversion) and facilitates unsupervised discovery of interpretable axes. Our results illustrate that defining the geometry of the GAN image manifold can serve as a general framework for understanding GANs.

What Makes Instance Discrimination Good for Transfer Learning?

- Nanxuan Zhao, Zhirong Wu, Rynson W. H. Lau, Stephen Lin
- abstract@open-review(Poster): Contrastive visual pretraining based on the instance discrimination pretext task has made significant progress. Notably, recent work on unsupervised pretraining has shown to surpass the supervised counterpart for finetuning downstream applications such as object detection and segmentation. It comes as a surprise that image annotations would be better left unused for transfer learning. In this work, we investigate the following problems: What makes instance discrimination pretraining good for transfer learning? What knowledge is actually learned and transferred from these models? From this understanding of instance discrimination, how can we better exploit human annotation labels for pretraining? Our findings are threefold. First, what truly matters for the transfer is low-level and mid-level representations, not high-level representations. Second, the intra-category invariance enforced by the traditional supervised model weakens transferability by increasing task misalignment. Finally, supervised pretraining can be strengthened by following an exemplar-based approach without explicit constraints among the instances within the same category.

Answering Complex Open-Domain Questions with Multi-Hop Dense Retrieval

- Wenhan Xiong, Xiang Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Scott Yih, Sebastian Riedel, Douwe Kiela, Barlas Oguz
- abstract@open-review(Poster): We propose a simple and efficient multi-hop dense retrieval approach for answering complex open-domain questions, which achieves state-of-the-art performance on two multi-hop datasets, HotpotQA and multi-evidence FEVER. Contrary to previous work, our method does not require access to any corpus-specific information, such as inter-document hyperlinks or human-annotated entity markers, and can be applied to any unstructured text corpus. Our system also yields a much better efficiency-accuracy trade-off, matching the best published accuracy on HotpotQA while being 10 times faster at inference time.

On Dyadic Fairness: Exploring and Mitigating Bias in Graph Connections

- Peizhao Li, Yifei Wang, Han Zhao, Pengyu Hong, Hongfu Liu
- abstract@open-review(Poster): Disparate impact has raised serious concerns in machine learning applications and its societal impacts. In response to the need of mitigating discrimination, fairness has been regarded as a crucial property in algorithmic design. In this work, we study the problem of disparate impact on graph-structured data. Specifically, we focus on dyadic fairness, which articulates a fairness concept that a predictive relationship between two instances should be independent of the sensitive attributes. Based on this, we theoretically relate the graph connections to dyadic fairness on link predictive scores in learning graph neural networks, and reveal that regulating weights on existing edges in a graph contributes to dyadic fairness conditionally. Subsequently, we propose our algorithm, \textbf{FairAdj}, to empirically learn a fair adjacency matrix with proper graph structural constraints for fair link prediction, and in the meanwhile preserve predictive accuracy as much as possible. Empirical validation demonstrates that our method delivers effective dyadic fairness in terms of various statistics, and at the same time enjoys a favorable fairness-utility tradeoff.

Spatio-Temporal Graph Scattering Transform

- Chao Pan, Siheng Chen, Antonio Ortega
- abstract@open-review(Poster): Although spatio-temporal graph neural networks have achieved great empirical success in handling multiple correlated time series, they may be impractical in some real-world scenarios due to a lack of sufficient high-quality training data. Furthermore, spatio-temporal graph neural networks lack theoretical interpretation. To address these issues, we put forth a novel mathematically designed framework to analyze spatio-temporal data. Our proposed spatio-temporal graph scattering transform (ST-GST) extends traditional scattering transform to the spatio-temporal domain. It performs iterative applications of spatio-temporal graph wavelets and nonlinear activation functions, which can be viewed as a forward pass of spatio-temporal graph convolutional networks without training. Since all the filter coefficients in ST-GST are mathematically designed, it is promising for the real-world scenarios with limited training data, and also allows for a theoretical analysis, which shows that the proposed ST-GST is stable to small perturbations of input signals and structures. Finally, our experiments show that i) ST-GST outperforms spatio-temporal graph convolutional networks by an increase of 35% in accuracy for MSR Action3D dataset; ii) it is better and computationally more efficient to design the transform based on separable spatio-temporal graphs than the joint ones; and iii) nonlinearity in ST-GST is critical to empirical performance.

Image Augmentation Is All You Need: Regularizing Deep Reinforcement Learning from Pixels

- Denis Yarats, Ilya Kostrikov, Rob Fergus
- abstract@open-review(Spotlight): We propose a simple data augmentation technique that can be applied to standard model-free reinforcement learning algorithms, enabling robust learning directly from pixels without the need for auxiliary losses or pre-training. The approach leverages input perturbations commonly used in computer vision tasks to transform input examples, as well as regularizing the value function and policy. Existing model-free approaches, such as Soft Actor-Critic (SAC), are not able to train deep networks effectively from image pixels. However, the addition of our augmentation method dramatically improves SAC's performance, enabling it to reach state-of-the-art performance on the DeepMind control suite, surpassing model-based (Hafner et al., 2019; Lee et al., 2019; Hafner et al., 2018) methods and recently proposed contrastive learning (Srinivas et al., 2020). Our approach, which we dub DrQ: Data-regularized Q, can be combined with any model-free reinforcement learning algorithm. We further demonstrate this by applying it to DQN and significantly improve its data-efficiency on the Atari 100k benchmark.

MODALS: Modality-agnostic Automated Data Augmentation in the Latent Space

- Tsz-Him Cheung, Dit-Yan Yeung
- abstract@open-review(Poster): Data augmentation is an efficient way to expand a training dataset by creating additional artificial data. While data augmentation is found to be effective in improving the generalization capabilities of models for various machine learning tasks, the underlying augmentation methods are usually manually designed and carefully evaluated for each data modality separately, like image processing functions for image data and word-replacing rules for text data. In this work, we propose an automated data augmentation approach called MODALS (Modality-agnostic Automated Data Augmentation in the Latent Space) to augment data for any modality in a generic way. MODALS exploits automated data augmentation to fine-tune four universal data transformation operations in the latent space to adapt the transform to data of different modalities. Through comprehensive experiments, we demonstrate the effectiveness of MODALS on multiple datasets for text, tabular, time-series and image modalities.

ALFWorld: Aligning Text and Embodied Environments for Interactive Learning

- Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, Matthew Hausknecht
- abstract@open-review(Poster): Given a simple request like Put a washed apple in the kitchen fridge, humans can reason in purely abstract terms by imagining action sequences and scoring their likelihood of success, prototypicality, and efficiency, all without moving a muscle. Once we see the kitchen in question, we can update our abstract plans to fit the scene. Embodied agents require the same abilities, but existing work does not yet provide the infrastructure necessary for both reasoning abstractly and executing concretely. We address this limitation by introducing ALFWorld, a simulator that

enables agents to learn abstract, text-based policies in TextWorld (Côté et al., 2018) and then execute goals from the ALFRED benchmark (Shridhar et al., 2020) in a rich visual environment. ALFWorld enables the creation of a new BUTLER agent whose abstract knowledge, learned in TextWorld, corresponds directly to concrete, visually grounded actions. In turn, as we demonstrate empirically, this fosters better agent generalization than training only in the visually grounded environment. BUTLER's simple, modular design factors the problem to allow researchers to focus on models for improving every piece of the pipeline (language understanding, planning, navigation, and visual scene understanding).

Latent Skill Planning for Exploration and Transfer

- Kevin Xie, Homanga Bharadhwaj, Danijar Hafner, Animesh Garg, Florian Shkurti
- abstract@open-review(Poster): To quickly solve new tasks in complex environments, intelligent agents need to build up reusable knowledge. For example, a learned world model captures knowledge about the environment that applies to new tasks. Similarly, skills capture general behaviors that can apply to new tasks. In this paper, we investigate how these two approaches can be integrated into a single reinforcement learning agent. Specifically, we leverage the idea of partial amortization for fast adaptation at test time. For this, actions are produced by a policy that is learned over time while the skills it conditions on are chosen using online planning. We demonstrate the benefits of our design decisions across a suite of challenging locomotion tasks and demonstrate improved sample efficiency in single tasks as well as in transfer from one task to another, as compared to competitive baselines. Videos are available at: https://sites.google.com/view/latent-skill-planning/

Evaluating the Disentanglement of Deep Generative Models through Manifold Topology

- Sharon Zhou, Eric Zelikman, Fred Lu, Andrew Y. Ng, Gunnar E. Carlsson, Stefano Ermon
- abstract@open-review(Poster): Learning disentangled representations is regarded as a fundamental task for improving the generalization, robustness, and interpretability of generative models. However, measuring disentanglement has been challenging and inconsistent, often dependent on an ad-hoc external model or specific to a certain dataset. To address this, we present a method for quantifying disentanglement that only uses the generative model, by measuring the topological similarity of conditional submanifolds in the learned representation. This method showcases both unsupervised and supervised variants. To illustrate the effectiveness and applicability of our method, we empirically evaluate several state-of-the-art models across multiple datasets. We find that our method ranks models similarly to existing methods. We make our code publicly available at https://github.com/stanfordmlgroup/disentanglement.

Combining Ensembles and Data Augmentation Can Harm Your Calibration

- Yeming Wen, Ghassen Jerfel, Rafael Muller, Michael W Dusenberry, Jasper Snoek, Balaji Lakshminarayanan, Dustin Tran
- abstract@open-review(Poster): Ensemble methods which average over multiple neural network predictions are a simple approach to improve a model's calibration and robustness. Similarly, data augmentation techniques, which encode prior information in the form of invariant feature transformations, are effective for improving calibration and robustness. In this paper, we show a surprising pathology: combining ensembles and data augmentation can harm model calibration. This leads to a trade-off in practice, whereby improved accuracy by combining the two techniques comes at the expense of calibration. On the other hand, selecting only one of the techniques ensures good uncertainty estimates at the expense of accuracy. We investigate this pathology and identify a compounding under-confidence among methods which marginalize over sets of weights and data augmentation techniques which soften labels. Finally, we propose a simple correction, achieving the best of both worlds with significant accuracy and calibration gains over using only ensembles or data augmentation individually. Applying the correction produces new state-of-the art in uncertainty calibration and robustness across CIFAR-10, CIFAR-100, and ImageNet.

Dynamic Tensor Rematerialization

- Marisa Kirisame, Steven Lyubomirsky, Altan Haan, Jennifer Brennan, Mike He, Jared Roesch, Tianqi Chen, Zachary Tatlock
- abstract@open-review(Spotlight): Checkpointing enables the training of deep learning models under restricted memory budgets by freeing intermediate activations from memory and recomputing them on demand. Current checkpointing techniques statically plan these recomputations offline and assume static computation graphs. We demonstrate that a simple online algorithm can achieve comparable performance by introducing Dynamic Tensor Rematerialization (DTR), a greedy online algorithm for checkpointing that is extensible and general, is parameterized by eviction policy, and supports dynamic models. We prove that DTR can train an N-layer linear feedforward network on an $\Omega(\sqrt{N})$ memory budget with only $\mathcal{O}(N)$ tensor operations. DTR closely matches the performance of optimal static checkpointing in simulated experiments. We incorporate a DTR prototype into PyTorch merely by interposing on tensor allocations and operator calls and collecting lightweight metadata on tensors.

CoCo: Controllable Counterfactuals for Evaluating Dialogue State Trackers

- SHIYANG LI, Semih Yavuz, Kazuma Hashimoto, Jia Li, Tong Niu, Nazneen Rajani, Xifeng Yan, Yingbo Zhou, Caiming Xiong
- abstract@open-review(Poster): Dialogue state trackers have made significant progress on benchmark datasets, but their generalization capability to novel and realistic scenarios beyond the held- out conversations is less understood. We propose controllable counterfactuals (COCO) to bridge this gap and evaluate dialogue state tracking (DST) models on novel scenarios, i.e., would the system successfully tackle the request if the user responded differently but still consistently with the dialogue flow? COCO leverages turn-level belief states as counterfactual conditionals to produce novel conversation scenarios in two steps: (i) counterfactual goal generation at turn- level by dropping and adding slots followed by replacing slot values, (ii) counterfactual conversation generation that is conditioned on (i) and consistent with the dialogue flow. Evaluating state-of-the-art DST models on MultiWOZ dataset with COCO-generated counterfactuals results in a significant performance drop of up to 30.8% (from 49.4% to 18.6%) in absolute joint goal accuracy. In comparison, widely used techniques like paraphrasing only affect the accuracy by at most 2%. Human evaluations show that COCO-generated conversations perfectly reflect the underlying user goal with more than 95% accuracy and are as human-like as the original conversations, further strengthening its reliability and promise to be adopted as part of the robustness evaluation of DST models.

Deep Neural Network Fingerprinting by Conferrable Adversarial Examples

- Nils Lukas, Yuxuan Zhang, Florian Kerschbaum
- abstract@open-review(Spotlight): In Machine Learning as a Service, a provider trains a deep neural network and gives many users access. The hosted (source) model is susceptible to model stealing attacks, where an adversary derives a surrogate model from API access to the source model. For post hoc detection of such attacks, the provider needs a robust method to determine whether a suspect model is a surrogate of their model. We propose a fingerprinting method for deep neural network classifiers that extracts a set of inputs from the source model so that only surrogates agree with the source model on the classification of such inputs. These inputs are a subclass of transferable adversarial examples which we call conferrable adversarial examples that exclusively transfer with a target label from a source model to its surrogates. We propose a new method to generate these conferrable adversarial examples. We present an extensive study on the irremovability of our fingerprint against fine-tuning, weight pruning, retraining, retraining with different architectures, three model extraction attacks from related work, transfer learning, adversarial training, and two new adaptive attacks. Our fingerprint is robust against distillation, related model extraction attacks, and even transfer learning when the attacker has no access to the model provider's dataset. Our fingerprint is the first method that reaches a ROC AUC of 1.0 in verifying surrogates, compared to a ROC AUC of 0.63 by previous fingerprints.

AdaGCN: Adaboosting Graph Convolutional Networks into Deep Models

- Ke Sun, Zhanxing Zhu, Zhouchen Lin
- abstract@open-review(Poster): The design of deep graph models still remains to be investigated and the crucial part is how to explore and exploit the knowledge from different hops of neighbors in an efficient way. In this paper, we propose a novel RNN-like deep graph neural network architecture by incorporating AdaBoost into the computation of network; and the proposed graph convolutional network called AdaGCN~(Adaboosting Graph Convolutional Network) has the ability to efficiently extract knowledge from high-order neighbors of current nodes and then integrates knowledge from different hops of neighbors into the network in an Adaboost way. Different from other graph neural networks that directly stack many graph convolution layers, AdaGCN shares the same base neural network architecture among all ``layers'' and is recursively optimized, which is similar to an RNN. Besides, We also theoretically established the connection between AdaGCN and existing graph convolutional methods, presenting the benefits of our proposal. Finally, extensive experiments demonstrate the consistent state-of-the-art prediction performance on graphs across different label rates and the computational advantage of our approach AdaGCN~\footnote{Code is available at \url{https://github.com/datake/AdaGCN}.}.

Learning What To Do by Simulating the Past

- David Lindner, Rohin Shah, Pieter Abbeel, Anca Dragan
- abstract@open-review(Poster): Since reward functions are hard to specify, recent work has focused on learning policies from human feedback. However, such approaches are impeded by the expense of acquiring such feedback. Recent work proposed that agents have access to a source of information that is effectively free: in any environment that humans have acted in, the state will already be optimized for human preferences, and thus an agent can extract information about what humans want from the state. Such learning is possible in principle, but requires simulating all possible past trajectories that could have led to the observed state. This is feasible in gridworlds, but how do we scale it to complex tasks? In this work, we show that by combining a learned feature encoder with learned inverse models, we can enable agents to simulate human actions backwards in time to infer what they must have done. The resulting algorithm is able to reproduce a specific skill in MuJoCo environments given a single state sampled from the optimal policy for that skill.

CcGAN: Continuous Conditional Generative Adversarial Networks for Image Generation

- Xin Ding, Yongwei Wang, Zuheng Xu, William J Welch, Z. Jane Wang
- abstract@open-review(Poster): This work proposes the continuous conditional generative adversarial network (CcGAN), the first generative model for image generation conditional on continuous, scalar conditions (termed regression labels). Existing conditional GANs (cGANs) are mainly designed for categorical conditions (e.g., class labels); conditioning on a continuous label is mathematically distinct and raises two fundamental problems: (P1) Since there may be very few (even zero) real images for some regression labels, minimizing existing empirical versions of cGAN losses (a.k.a. empirical cGAN losses) often fails in practice; (P2) Since regression labels are scalar and infinitely many, conventional label input methods (e.g., combining a hidden map of the generator/discriminator with a one-hot encoded label) are not applicable. The proposed CcGAN solves the above problems, respectively, by (S1) reformulating existing empirical cGAN losses to be appropriate for the continuous scenario; and (S2) proposing a novel method to incorporate regression labels into the generator and the discriminator. The reformulation in (S1) leads to two novel empirical discriminator losses, termed the hard vicinal discriminator loss (HVDL) and the soft vicinal discriminator loss (SVDL) respectively, and a novel empirical generator loss. The error bounds of a discriminator trained with HVDL and SVDL are derived under mild assumptions in this work. A new benchmark dataset, RC-49, is also proposed for generative image modeling conditional on regression labels. Our experiments on the Circular 2-D Gaussians, RC-49, and UTKFace datasets show that CcGAN is able to generate diverse, high-quality samples from the image distribution conditional on a given regression label. Moreover, in these experiments, CcGAN substantially outperforms cGAN both visually and quantitatively.

ANOCE: Analysis of Causal Effects with Multiple Mediators via Constrained Structural Learning

- Hengrui Cai, Rui Song, Wenbin Lu
- abstract@open-review(Poster): In the era of causal revolution, identifying the causal effect of an exposure on the outcome of interest is an important problem in many areas, such as epidemics, medicine, genetics, and economics. Under a general causal graph, the exposure may have a direct effect on the outcome and also an indirect effect regulated by a set of mediators. An analysis of causal effects that interprets the causal mechanism contributed through mediators is hence challenging but on demand. To the best of our knowledge, there are no feasible algorithms that give an exact decomposition of the indirect effect on the level of individual mediators, due to common interaction among mediators in the complex graph. In this paper, we establish a new statistical framework to comprehensively characterize causal effects with multiple mediators, namely, ANalysis Of Causal Effects (ANOCE), with a newly introduced definition of the mediator effect, under the linear structure equation model. We further propose a constrained causal structure learning method by incorporating a novel identification constraint that specifies the temporal causal relationship of variables. The proposed algorithm is applied to investigate the causal effects of 2020 Hubei lockdowns on reducing the spread of the coronavirus in Chinese major cities out of Hubei.

Direction Matters: On the Implicit Bias of Stochastic Gradient Descent with Moderate Learning Rate

- Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu
- abstract@open-review(Poster): Understanding the algorithmic bias of stochastic gradient descent (SGD) is one of the key challenges in modern machine learning and deep learning theory. Most of the existing works, however, focus on very small or even infinitesimal learning rate regime, and fail to cover practical scenarios where the learning rate is moderate and annealing. In this paper, we make an initial attempt to characterize the particular regularization effect of SGD in the moderate learning rate regime by studying its behavior for optimizing an overparameterized linear regression problem. In this case, SGD and GD are known to converge to the unique minimum-norm solution; however, with the moderate and annealing learning rate, we show that they exhibit different directional bias: SGD converges along the large eigenvalue directions of the data matrix, while GD goes after the small eigenvalue directions. Furthermore, we show that such directional bias does matter when early stopping is adopted, where the SGD output is nearly optimal but the GD output is suboptimal. Finally, our theory explains several folk arts in practice used for SGD hyperparameter tuning, such as (1) linearly scaling the initial learning rate with batch size; and (2) overrunning SGD with high learning rate even when the loss stops decreasing.

Robust early-learning: Hindering the memorization of noisy labels

- Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, Yi Chang
- abstract@open-review(Poster): The \textit{memorization effects} of deep networks show that they will first memorize training data with clean labels and then those with noisy labels. The \textit{early stopping} method therefore can be exploited for learning with noisy labels. However, the side effect brought by noisy labels will influence the memorization of clean labels before early stopping. In this paper, motivated by the \textit{lottery ticket hypothesis} which shows that only partial parameters are important for generalization, we find that only partial parameters are important for fitting clean labels and generalize well, which we term as \textit{critical parameters}; while the other parameters tend to fit noisy labels and cannot generalize well, which we term as \textit{non-critical parameters}. Based on this, we propose \textit{robust early-learning} to reduce the side effect of noisy labels before early stopping and thus enhance the memorization of clean labels. Specifically, in each iteration, we divide all parameters into the critical and non-critical ones, and then perform different update rules for different types of parameters. Extensive experiments on benchmark-simulated and real-world label-noise datasets demonstrate the superiority of the proposed method over the state-of-the-art label-noise learning methods.

SMiRL: Surprise Minimizing Reinforcement Learning in Unstable Environments

- Glen Berseth, Daniel Geng, Coline Manon Devin, Nicholas Rhinehart, Chelsea Finn, Dinesh Jayaraman, Sergey Levine

- abstract@open-review(Oral): Every living organism struggles against disruptive environmental forces to carve out and maintain an orderly niche. We propose that such a struggle to achieve and preserve order might offer a principle for the emergence of useful behaviors in artificial agents. We formalize this idea into an unsupervised reinforcement learning method called surprise minimizing reinforcement learning (SMiRL). SMiRL alternates between learning a density model to evaluate the surprise of a stimulus, and improving the policy to seek more predictable stimuli. The policy seeks out stable and repeatable situations that counteract the environment's prevailing sources of entropy. This might include avoiding other hostile agents, or finding a stable, balanced pose for a bipedal robot in the face of disturbance forces. We demonstrate that our surprise minimizing agents can successfully play Tetris, Doom, control a humanoid to avoid falls, and navigate to escape enemies in a maze without any task-specific reward supervision. We further show that SMiRL can be used together with standard task rewards to accelerate reward-driven learning.

A Design Space Study for LISTA and Beyond

- Tianjian Meng, Xiaohan Chen, Yifan Jiang, Zhangyang Wang
- abstract@open-review(Poster): In recent years, great success has been witnessed in building problem-specific deep networks from unrolling iterative algorithms, for solving inverse problems and beyond. Unrolling is believed to incorporate the model-based prior with the learning capacity of deep learning. This paper revisits \textit{the role of unrolling as a design approach for deep networks}: to what extent its resulting special architecture is superior, and can we find better? Using LISTA for sparse recovery as a representative example, we conduct the first thorough \textit{design space study} for the unrolled models. Among all possible variations, we focus on extensively varying the connectivity patterns and neuron types, leading to a gigantic design space arising from LISTA. To efficiently explore this space and identify top performers, we leverage the emerging tool of neural architecture search (NAS). We carefully examine the searched top architectures in a number of settings, and are able to discover networks that consistently better than LISTA. We further present more visualization and analysis to ``open the black box", and find that the searched top architectures demonstrate highly consistent and potentially transferable patterns. We hope our study to spark more reflections and explorations on how to better mingle model-based optimization prior and data-driven learning.

PAC Confidence Predictions for Deep Neural Network Classifiers

- Sangdon Park, Shuo Li, Insup Lee, Osbert Bastani
- abstract@open-review(Poster): A key challenge for deploying deep neural networks (DNNs) in safety critical settings is the need to provide rigorous ways to quantify their uncertainty. In this paper, we propose a novel algorithm for constructing predicted classification confidences for DNNs that comes with provable correctness guarantees. Our approach uses Clopper-Pearson confidence intervals for the Binomial distribution in conjunction with the histogram binning approach to calibrated prediction. In addition, we demonstrate how our predicted confidences can be used to enable downstream guarantees in two settings: (i) fast DNN inference, where we demonstrate how to compose a fast but inaccurate DNN with an accurate but slow DNN in a rigorous way to improve performance without sacrificing accuracy, and (ii) safe planning, where we guarantee safety when using a DNN to predict whether a given action is safe based on visual observations. In our experiments, we demonstrate that our approach can be used to provide guarantees for state-of-the-art DNNs.

X2T: Training an X-to-Text Typing Interface with Online Learning from User Feedback

- Jensen Gao, Siddharth Reddy, Glen Berseth, Nicholas Hardy, Nikhilesh Natraj, Karunesh Ganguly, Anca Dragan, Sergey Levine
- abstract@open-review(Poster): We aim to help users communicate their intent to machines using flexible, adaptive interfaces that translate arbitrary user input into desired actions. In this work, we focus on assistive typing applications in which a user cannot operate a keyboard, but can instead supply other inputs, such as webcam images that capture eye gaze or neural activity measured by a brain implant. Standard methods train a model on a fixed dataset of user inputs, then deploy a static interface that does not learn from its mistakes; in part, because extracting an error signal from user behavior can be challenging. We investigate a simple idea that would enable such interfaces to improve over time, with minimal additional effort from the user: online learning from user feedback on the accuracy of the interface's actions. In the typing domain, we leverage backspaces as feedback that the interface did not perform the desired action. We propose an algorithm called x-to-text (X2T) that trains a predictive model of this feedback signal, and uses this model to fine-tune any existing, default interface for translating user input into actions that select words or characters. We evaluate X2T through a small-scale online user study with 12 participants who type sentences by gazing at their desired words, a large-scale observational study on handwriting samples from 60 users, and a pilot study with one participant using an electrocorticography-based brain-computer interface. The results show that X2T learns to outperform a non-adaptive default interface, stimulates user co-adaptation to the interface, personalizes the interface to individual users, and can leverage offline data collected from the default interface to improve its initial performance and accelerate online learning.

Discrete Graph Structure Learning for Forecasting Multiple Time Series

- Chao Shang, Jie Chen, Jinbo Bi
- abstract@open-review(Poster): Time series forecasting is an extensively studied subject in statistics, economics, and computer science. Exploration of the correlation and causation among the variables in a multivariate time series shows promise in enhancing the performance of a time series model. When using deep neural networks as forecasting models, we hypothesize that exploiting the pairwise information among multiple (multivariate) time series also improves their forecast. If an explicit graph structure is known, graph neural networks (GNNs) have been demonstrated as powerful tools to exploit the structure. In this work, we propose learning the structure simultaneously with the GNN if the graph is unknown. We cast the problem as learning a probabilistic graph model through optimizing the mean performance over the graph distribution. The distribution is parameterized by a neural network so that discrete graphs can be sampled differentiably through reparameterization. Empirical evaluations show that our method is simpler, more efficient, and better performing than a recently proposed bilevel learning approach for graph structure learning, as well as a broad array of forecasting models, either deep or non-deep learning based, and graph or non-graph based.

Task-Agnostic Morphology Evolution

- Donald Joseph Hejna III, Pieter Abbeel, Lerrel Pinto
- abstract@open-review(Poster): Deep reinforcement learning primarily focuses on learning behavior, usually overlooking the fact that an agent's function is largely determined by form. So, how should one go about finding a morphology fit for solving tasks in a given environment? Current approaches that co-adapt morphology and behavior use a specific task's reward as a signal for morphology optimization. However, this often requires expensive policy optimization and results in task-dependent morphologies that are not built to generalize. In this work, we propose a new approach, Task-Agnostic Morphology Evolution (TAME), to alleviate both of these issues. Without any task or reward specification, TAME evolves morphologies by only applying randomly sampled action primitives on a population of agents. This is accomplished using an information-theoretic objective that efficiently ranks agents by their ability to reach diverse states in the environment and the causality of their actions. Finally, we empirically demonstrate that across 2D, 3D, and manipulation environments TAME can evolve morphologies that match the multi-task performance of those learned with task supervised algorithms. Our code and videos can be found at https://sites.google.com/view/task-agnostic-evolution .

Deep Equals Shallow for ReLU Networks in Kernel Regimes

- Alberto Bietti, Francis Bach
- abstract@open-review(Poster): Deep networks are often considered to be more expressive than shallow ones in terms of approximation. Indeed, certain functions can be approximated by deep networks provably more efficiently than by shallow ones, however, no tractable algorithms are known for learning such deep models. Separately, a recent line of work has shown that deep networks trained with gradient descent may behave like (tractable) kernel

methods in a certain over-parameterized regime, where the kernel is determined by the architecture and initialization, and this paper focuses on approximation for such kernels. We show that for ReLU activations, the kernels derived from deep fully-connected networks have essentially the same approximation properties as their shallow two-layer counterpart, namely the same eigenvalue decay for the corresponding integral operator. This highlights the limitations of the kernel framework for understanding the benefits of such deep architectures. Our main theoretical result relies on characterizing such eigenvalue decays through differentiability properties of the kernel function, which also easily applies to the study of other kernels defined on the sphere.

Loss Function Discovery for Object Detection via Convergence-Simulation Driven Search

- Peidong Liu, Gengwei Zhang, Bochao Wang, Hang Xu, Xiaodan Liang, Yong Jiang, Zhenguo Li
- abstract@open-review(Poster): Designing proper loss functions for vision tasks has been a long-standing research direction to advance the capability of existing models. For object detection, the well-established classification and regression loss functions have been carefully designed by considering diverse learning challenges (e.g. class imbalance, hard negative samples, and scale variances). Inspired by the recent progress in network architecture search, it is interesting to explore the possibility of discovering new loss function formulations via directly searching the primitive operation combinations. So that the learned losses not only fit for diverse object detection challenges to alleviate huge human efforts, but also have better alignment with evaluation metric and good mathematical convergence property. Beyond the previous auto-loss works on face recognition and image classification, our work makes the first attempt to discover new loss functions for the challenging object detection from primitive operation levels and finds the searched losses are insightful. We propose an effective convergence-simulation driven evolutionary search algorithm, called CSE-Autoloss, for speeding up the search progress by regularizing the mathematical rationality of loss candidates via two progressive convergence simulation modules: convergence property verification and model optimization simulation. CSE-Autoloss involves the search space (i.e. 21 mathematical operators, 3 constant-type inputs, and 3 variable-type inputs) that cover a wide range of the possible variants of existing losses and discovers best-searched loss function combination within a short time (around 1.5 wall-clock days with 20x speedup in comparison to the vanilla evolutionary algorithm). We conduct extensive evaluations of loss function search on popular detectors and validate the good generalization capability of searched losses across diverse architectures and various datasets. Our experiments show that the best-discovered loss function combinations outperform default combinations (Cross-entropy/Focal loss for classification and L1 loss for regression) by 1.1% and 0.8% in terms of mAP for two-stage and one-stage detectors on COCO respectively. Our searched losses are available at https://github.com/PerdonLiu/CSE-Autoloss.

Effective Distributed Learning with Random Features: Improved Bounds and Algorithms

- Yong Liu, Jiankun Liu, Shuqiang Wang
- abstract@open-review(Poster): In this paper, we study the statistical properties of distributed kernel ridge regression together with random features (DKRR-RF), and obtain optimal generalization bounds under the basic setting, which can substantially relax the restriction on the number of local machines in the existing state-of-art bounds. Specifically, we first show that the simple combination of divide-and-conquer technique and random features can achieve the same statistical accuracy as the exact KRR in expectation requiring only $\mathcal{O}(|\mathcal{D}|)$ memory and $\mathcal{O}(|\mathcal{D}|^{1.5})$ time. Then, beyond the generalization bounds in expectation that demonstrate the average information for multiple trails, we derive generalization bounds in probability to capture the learning performance for a single trail. Finally, we propose an effective communication strategy to further improve the performance of DKRR-RF, and validate the theoretical bounds via numerical experiments.

On InstaHide, Phase Retrieval, and Sparse Matrix Factorization

- Sitan Chen, Xiaoxiao Li, Zhao Song, Danyang Zhuo
- abstract@open-review(Poster): In this work, we examine the security of InstaHide, a scheme recently proposed by \cite{hsla20} for preserving the security of private datasets in the context of distributed learning. To generate a synthetic training example to be shared among the distributed learners, InstaHide takes a convex combination of private feature vectors and randomly flips the sign of each entry of the resulting vector with probability 1/2. A salient question is whether this scheme is secure in any provable sense, perhaps under a plausible complexity-theoretic assumption.

The answer to this turns out to be quite subtle and closely related to the average-case complexity of a multi-task, missing-data version of the classic problem of phase retrieval that is interesting in its own right. Motivated by this connection, under the standard distributional assumption that the public/private feature vectors are isotropic Gaussian, we design an algorithm that can actually recover a private vector using only the public vectors and a sequence of synthetic vectors generated by InstaHide.

Anchor & Transform: Learning Sparse Embeddings for Large Vocabularies

- Paul Pu Liang, Manzil Zaheer, Yuan Wang, Amr Ahmed
- abstract@open-review(Poster): Learning continuous representations of discrete objects such as text, users, movies, and URLs lies at the heart of many applications including language and user modeling. When using discrete objects as input to neural networks, we often ignore the underlying structures (e.g., natural groupings and similarities) and embed the objects independently into individual vectors. As a result, existing methods do not scale to large vocabulary sizes. In this paper, we design a simple and efficient embedding algorithm that learns a small set of anchor embeddings and a sparse transformation matrix. We call our method Anchor & Transform (ANT) as the embeddings of discrete objects are a sparse linear combination of the anchors, weighted according to the transformation matrix. ANT is scalable, flexible, and end-to-end trainable. We further provide a statistical interpretation of our algorithm as a Bayesian nonparametric prior for embeddings that encourages sparsity and leverages natural groupings among objects. By deriving an approximate inference algorithm based on Small Variance Asymptotics, we obtain a natural extension that automatically learns the optimal number of anchors instead of having to tune it as a hyperparameter. On text classification, language modeling, and movie recommendation benchmarks, we show that ANT is particularly suitable for large vocabulary sizes and demonstrates stronger performance with fewer parameters (up to 40x compression) as compared to existing compression baselines.

Achieving Linear Speedup with Partial Worker Participation in Non-IID Federated Learning

- Haibo Yang, Minghong Fang, Jia Liu
- abstract@open-review(Poster): Federated learning (FL) is a distributed machine learning architecture that leverages a large number of workers to jointly learn a model with decentralized data. FL has received increasing attention in recent years thanks to its data privacy protection, communication efficiency and a linear speedup for convergence in training (i.e., convergence performance increases linearly with respect to the number of workers). However, existing studies on linear speedup for convergence are only limited to the assumptions of i.i.d. datasets across workers and/or full worker participation, both of which rarely hold in practice. So far, it remains an open question whether or not the linear speedup for convergence is achievable under non-i.i.d. datasets with partial worker participation in FL. In this paper, we show that the answer is affirmative. Specifically, we show that the federated averaging (FedAvg) algorithm (with two-sided learning rates) on non-i.i.d. datasets in non-convex settings achieves a convergence rate $\mathcal{O}(\frac{1}{\sqrt{mKT}} + \frac{1}{T})$ for full worker participation and a convergence rate $\mathcal{O}(\frac{\sqrt{K}}{\sqrt{nT}} + \frac{1}{T})$ for partial worker participation, where K is the number of local steps, T is the number of total communication rounds, m is the total worker number and n is the worker number in one communication round if for partial worker participation. Our results also reveal that the local steps in FL could help the convergence and show that the maximum number of local steps can be improved to T/m in full worker participation. We conduct extensive experiments on MNIST and CIFAR-10 to verify our theoretical results.

Deep Neural Tangent Kernel and Laplace Kernel Have the Same RKHS

- Lin Chen, Sheng Xu
- abstract@open-review(Poster): We prove that the reproducing kernel Hilbert spaces (RKHS) of a deep neural tangent kernel and the Laplace kernel include the same set of functions, when both kernels are restricted to the sphere $\mathbb{S}^{d-1}$. Additionally, we prove that the exponential power kernel with a smaller power (making the kernel less smooth) leads to a larger RKHS, when it is restricted to the sphere $\mathbb{S}^{d-1}$ and when it is defined on the entire $\mathbb{R}^d$.

Fast convergence of stochastic subgradient method under interpolation

- Huang Fang, Zhenan Fan, Michael Friedlander
- abstract@open-review(Poster): This paper studies the behaviour of the stochastic subgradient descent (SSGD) method applied to over-parameterized nonsmooth optimization problems that satisfy an interpolation condition. By leveraging the composite structure of the empirical risk minimization problems, we prove that SSGD converges, respectively, with rates $O(1/\epsilon)$ and $O(\log(1/\epsilon))$ for convex and strongly-convex objectives when interpolation holds. These rates coincide with established rates for the stochastic gradient descent (SGD) method applied to smooth problems that also satisfy an interpolation condition. Our analysis provides a partial explanation for the empirical observation that sometimes SGD and SSGD behave similarly for training smooth and nonsmooth machine learning models. We also prove that the rate $O(1/\epsilon)$ is optimal for the subgradient method in the convex and interpolation setting.

Generating Adversarial Computer Programs using Optimized Obfuscations

- Shashank Srikant, Sijia Liu, Tamara Mitrovska, Shiyu Chang, Quanfu Fan, Gaoyuan Zhang, Una-May O'Reilly
- abstract@open-review(Poster): Machine learning (ML) models that learn and predict properties of computer programs are increasingly being adopted and deployed. These models have demonstrated success in applications such as auto-completing code, summarizing large programs, and detecting bugs and malware in programs. In this work, we investigate principled ways to adversarially perturb a computer program to fool such learned models, and thus determine their adversarial robustness. We use program obfuscations, which have conventionally been used to avoid attempts at reverse engineering programs, as adversarial perturbations. These perturbations modify programs in ways that do not alter their functionality but can be crafted to deceive an ML model when making a decision. We provide a general formulation for an adversarial program that allows applying multiple obfuscation transformations to a program in any language. We develop first-order optimization algorithms to efficiently determine two key aspects -- which parts of the program to transform, and what transformations to use. We show that it is important to optimize both these aspects to generate the best adversarially perturbed program. Due to the discrete nature of this problem, we also propose using randomized smoothing to improve the attack loss landscape to ease optimization. We evaluate our work on Python and Java programs on the problem of program summarization. We show that our best attack proposal achieves a 52% improvement over a state-of-the-art attack generation approach for programs trained on a \textsc{seq2seq} model. We further show that our formulation is better at training models that are robust to adversarial attacks.

C-Learning: Learning to Achieve Goals via Recursive Classification

- Benjamin Eysenbach, Ruslan Salakhutdinov, Sergey Levine
- abstract@open-review(Poster): We study the problem of predicting and controlling the future state distribution of an autonomous agent. This problem, which can be viewed as a reframing of goal-conditioned reinforcement learning (RL), is centered around learning a conditional probability density function over future states. Instead of directly estimating this density function, we indirectly estimate this density function by training a classifier to predict whether an observation comes from the future. Via Bayes' rule, predictions from our classifier can be transformed into predictions over future states. Importantly, an off-policy variant of our algorithm allows us to predict the future state distribution of a new policy, without collecting new experience. This variant allows us to optimize functionals of a policy's future state distribution, such as the density of reaching a particular goal state. While conceptually similar to Q-learning, our work lays a principled foundation for goal-conditioned RL as density estimation, providing justification for goal-conditioned methods used in prior work. This foundation makes hypotheses about Q-learning, including the optimal goal-sampling ratio, which we confirm experimentally. Moreover, our proposed method is competitive with prior goal-conditioned RL methods.

Off-Dynamics Reinforcement Learning: Training for Transfer with Domain Classifiers

- Benjamin Eysenbach, Shreyas Chaudhari, Swapnil Asawa, Sergey Levine, Ruslan Salakhutdinov
- abstract@open-review(Poster): We propose a simple, practical, and intuitive approach for domain adaptation in reinforcement learning. Our approach stems from the idea that the agent's experience in the source domain should look similar to its experience in the target domain. Building off of a probabilistic view of RL, we achieve this goal by compensating for the difference in dynamics by modifying the reward function. This modified reward function is simple to estimate by learning auxiliary classifiers that distinguish source-domain transitions from target-domain transitions. Intuitively, the agent is penalized for transitions that would indicate that the agent is interacting with the source domain, rather than the target domain. Formally, we prove that applying our method in the source domain is guaranteed to obtain a near-optimal policy for the target domain, provided that the source and target domains satisfy a lightweight assumption. Our approach is applicable to domains with continuous states and actions and does not require learning an explicit model of the dynamics. On discrete and continuous control tasks, we illustrate the mechanics of our approach and demonstrate its scalability to high-dimensional~tasks.

Learning to Reach Goals via Iterated Supervised Learning

- Dibya Ghosh, Abhishek Gupta, Ashwin Reddy, Justin Fu, Coline Manon Devin, Benjamin Eysenbach, Sergey Levine
- abstract@open-review(Oral): Current reinforcement learning (RL) algorithms can be brittle and difficult to use, especially when learning goal-reaching behaviors from sparse rewards. Although supervised imitation learning provides a simple and stable alternative, it requires access to demonstrations from a human supervisor. In this paper, we study RL algorithms that use imitation learning to acquire goal reaching policies from scratch, without the need for expert demonstrations or a value function. In lieu of demonstrations, we leverage the property that any trajectory is a successful demonstration for reaching the final state in that same trajectory. We propose a simple algorithm in which an agent continually relabels and imitates the trajectories it generates to progressively learn goal-reaching behaviors from scratch. Each iteration, the agent collects new trajectories using the latest policy, and maximizes the likelihood of the actions along these trajectories under the goal that was actually reached, so as to improve the policy. We formally show that this iterated supervised learning procedure optimizes a bound on the RL objective, derive performance bounds of the learned policy, and empirically demonstrate improved goal-reaching performance and robustness over current RL algorithms in several benchmark tasks.

Model-Based Visual Planning with Self-Supervised Functional Distances

- Stephen Tian, Suraj Nair, Frederik Ebert, Sudeep Dasari, Benjamin Eysenbach, Chelsea Finn, Sergey Levine
- abstract@open-review(Spotlight): A generalist robot must be able to complete a variety of tasks in its environment. One appealing way to specify each task is in terms of a goal observation. However, learning goal-reaching policies with reinforcement learning remains a challenging problem, particularly when hand-engineered reward functions are not available. Learned dynamics models are a promising approach for learning about the environment without rewards or task-directed data, but planning to reach goals with such a model requires a notion of functional similarity between observations and goal

states. We present a self-supervised method for model-based visual goal reaching, which uses both a visual dynamics model as well as a dynamical distance function learned using model-free reinforcement learning. Our approach learns entirely using offline, unlabeled data, making it practical to scale to large and diverse datasets. In our experiments, we find that our method can successfully learn models that perform a variety of tasks at test-time, moving objects amid distractors with a simulated robotic arm and even learning to open and close a drawer using a real-world robot. In comparisons, we find that this approach substantially outperforms both model-free and model-based prior methods.

Mathematical Reasoning via Self-supervised Skip-tree Training

- Markus Norman Rabe, Dennis Lee, Kshitij Bansal, Christian Szegedy
- abstract@open-review(Spotlight): We demonstrate that self-supervised language modeling applied to mathematical formulas enables logical reasoning. To measure the logical reasoning abilities of language models, we formulate several evaluation (downstream) tasks, such as inferring types, suggesting missing assumptions and completing equalities. For training language models for formal mathematics, we propose a novel skip-tree task. We find that models trained on the skip-tree task show surprisingly strong mathematical reasoning abilities, and outperform models trained on standard skip-sequence tasks. We also analyze the models' ability to formulate new conjectures by measuring how often the predictions are provable and useful in other proofs.

DeepAveragers: Offline Reinforcement Learning By Solving Derived Non-Parametric MDPs

- Aayam Kumar Shrestha, Stefan Lee, Prasad Tadepalli, Alan Fern
- abstract@open-review(Spotlight): We study an approach to offline reinforcement learning (RL) based on optimally solving finitely-represented MDPs derived from a static dataset of experience. This approach can be applied on top of any learned representation and has the potential to easily support multiple solution objectives as well as zero-shot adjustment to changing environments and goals. Our main contribution is to introduce the Deep Averagers with Costs MDP (DAC-MDP) and to investigate its solutions for offline RL. DAC-MDPs are a non-parametric model that can leverage deep representations and account for limited data by introducing costs for exploiting under-represented parts of the model. In theory, we show conditions that allow for lower-bounding the performance of DAC-MDP solutions. We also investigate the empirical behavior in a number of environments, including those with image-based observations. Overall, the experiments demonstrate that the framework can work in practice and scale to large complex offline RL problems.

Multi-resolution modeling of a discrete stochastic process identifies causes of cancer

- Adam Uri Yaari, Maxwell Sherman, Oliver Clarke Priebe, Po-Ru Loh, Boris Katz, Andrei Barbu, Bonnie Berger
- abstract@open-review(Poster): Detection of cancer-causing mutations within the vast and mostly unexplored human genome is a major challenge. Doing so requires modeling the background mutation rate, a highly non-stationary stochastic process, across regions of interest varying in size from one to millions of positions. Here, we present the split-Poisson-Gamma (SPG) distribution, an extension of the classical Poisson-Gamma formulation, to model a discrete stochastic process at multiple resolutions. We demonstrate that the probability model has a closed-form posterior, enabling efficient and accurate linear-time prediction over any length scale after the parameters of the model have been inferred a single time. We apply our framework to model mutation rates in tumors and show that model parameters can be accurately inferred from high-dimensional epigenetic data using a convolutional neural network, Gaussian process, and maximum-likelihood estimation. Our method is both more accurate and more efficient than existing models over a large range of length scales. We demonstrate the usefulness of multi-resolution modeling by detecting genomic elements that drive tumor emergence and are of vastly differing sizes.

On the Theory of Implicit Deep Learning: Global Convergence with Implicit Layers

- Kenji Kawaguchi
- abstract@open-review(Spotlight): A deep equilibrium model uses implicit layers, which are implicitly defined through an equilibrium point of an infinite sequence of computation. It avoids any explicit computation of the infinite sequence by finding an equilibrium point directly via root-finding and by computing gradients via implicit differentiation. In this paper, we analyze the gradient dynamics of deep equilibrium models with nonlinearity only on weight matrices and non-convex objective functions of weights for regression and classification. Despite non-convexity, convergence to global optimum at a linear rate is guaranteed without any assumption on the width of the models, allowing the width to be smaller than the output dimension and the number of data points. Moreover, we prove a relation between the gradient dynamics of the deep implicit layer and the dynamics of trust region Newton method of a shallow explicit layer. This mathematically proven relation along with our numerical observation suggests the importance of understanding implicit bias of implicit layers and an open problem on the topic. Our proofs deal with implicit layers, weight tying and nonlinearity on weights, and differ from those in the related literature.

On the Critical Role of Conventions in Adaptive Human-AI Collaboration

- Andy Shih, Arjun Sawhney, Jovana Kondic, Stefano Ermon, Dorsa Sadigh
- abstract@open-review(Poster): Humans can quickly adapt to new partners in collaborative tasks (e.g. playing basketball), because they understand which fundamental skills of the task (e.g. how to dribble, how to shoot) carry over across new partners. Humans can also quickly adapt to similar tasks with the same partners by carrying over conventions that they have developed (e.g. raising hand signals pass the ball), without learning to coordinate from scratch. To collaborate seamlessly with humans, AI agents should adapt quickly to new partners and new tasks as well. However, current approaches have not attempted to distinguish between the complexities intrinsic to a task and the conventions used by a partner, and more generally there has been little focus on leveraging conventions for adapting to new settings. In this work, we propose a learning framework that teases apart rule-dependent representation from convention-dependent representation in a principled way. We show that, under some assumptions, our rule-dependent representation is a sufficient statistic of the distribution over best-response strategies across partners. Using this separation of representations, our agents are able to adapt quickly to new partners, and to coordinate with old partners on new tasks in a zero-shot manner. We experimentally validate our approach on three collaborative tasks varying in complexity: a contextual multi-armed bandit, a block placing task, and the card game Hanabi.

WaveGrad: Estimating Gradients for Waveform Generation

- Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, William Chan
- abstract@open-review(Poster): This paper introduces WaveGrad, a conditional model for waveform generation which estimates gradients of the data density. The model is built on prior work on score matching and diffusion probabilistic models. It starts from a Gaussian white noise signal and iteratively refines the signal via a gradient-based sampler conditioned on the mel-spectrogram. WaveGrad offers a natural way to trade inference speed for sample quality by adjusting the number of refinement steps, and bridges the gap between non-autoregressive and autoregressive models in terms of audio quality. We find that it can generate high fidelity audio samples using as few as six iterations. Experiments reveal WaveGrad to generate high fidelity audio, outperforming adversarial non-autoregressive baselines and matching a strong likelihood-based autoregressive baseline using fewer sequential operations. Audio samples are available at https://wavegrad.github.io/.

BUSTLE: Bottom-Up Program Synthesis Through Learning-Guided Exploration

- Augustus Odena, Kensen Shi, David Bieber, Rishabh Singh, Charles Sutton, Hanjun Dai

- abstract@open-review(Spotlight): Program synthesis is challenging largely because of the difficulty of search in a large space of programs. Human programmers routinely tackle the task of writing complex programs by writing sub-programs and then analyzing their intermediate results to compose them in appropriate ways. Motivated by this intuition, we present a new synthesis approach that leverages learning to guide a bottom-up search over programs. In particular, we train a model to prioritize compositions of intermediate values during search conditioned on a given set of input-output examples. This is a powerful combination because of several emergent properties. First, in bottom-up search, intermediate programs can be executed, providing semantic information to the neural network. Second, given the concrete values from those executions, we can exploit rich features based on recent work on property signatures. Finally, bottom-up search allows the system substantial flexibility in what order to generate the solution, allowing the synthesizer to build up a program from multiple smaller sub-programs. Overall, our empirical evaluation finds that the combination of learning and bottom-up search is remarkably effective, even with simple supervised learning approaches. We demonstrate the effectiveness of our technique on two datasets, one from the SyGuS competition and one of our own creation.

A Critique of Self-Expressive Deep Subspace Clustering

- Benjamin David Haeffele, Chong You, Rene Vidal
- abstract@open-review(Poster): Subspace clustering is an unsupervised clustering technique designed to cluster data that is supported on a union of linear subspaces, with each subspace defining a cluster with dimension lower than the ambient space. Many existing formulations for this problem are based on exploiting the self-expressive property of linear subspaces, where any point within a subspace can be represented as linear combination of other points within the subspace. To extend this approach to data supported on a union of non-linear manifolds, numerous studies have proposed learning an embedding of the original data using a neural network which is regularized by a self-expressive loss function on the data in the embedded space to encourage a union of linear subspaces prior on the data in the embedded space. Here we show that there are a number of potential flaws with this approach which have not been adequately addressed in prior work. In particular, we show the model formulation is often ill-posed in that it can lead to a degenerate embedding of the data, which need not correspond to a union of subspaces at all and is poorly suited for clustering. We validate our theoretical results experimentally and also repeat prior experiments reported in the literature, where we conclude that a significant portion of the previously claimed performance benefits can be attributed to an ad-hoc post processing step rather than the deep subspace clustering model.

Explaining by Imitating: Understanding Decisions by Interpretable Policy Learning

- Alihan Hüyük, Daniel Jarrett, Cem Tekin, Mihaela van der Schaar
- abstract@open-review(Poster): Understanding human behavior from observed data is critical for transparency and accountability in decision-making. Consider real-world settings such as healthcare, in which modeling a decision-maker's policy is challenging—with no access to underlying states, no knowledge of environment dynamics, and no allowance for live experimentation. We desire learning a data-driven representation of decision- making behavior that (1) inheres transparency by design, (2) accommodates partial observability, and (3) operates completely offline. To satisfy these key criteria, we propose a novel model-based Bayesian method for interpretable policy learning ("Interpole") that jointly estimates an agent's (possibly biased) belief-update process together with their (possibly suboptimal) belief-action mapping. Through experiments on both simulated and real-world data for the problem of Alzheimer's disease diagnosis, we illustrate the potential of our approach as an investigative device for auditing, quantifying, and understanding human decision-making behavior.

For self-supervised learning, Rationality implies generalization, provably

- Yamini Bansal, Gal Kaplun, Boaz Barak
- abstract@open-review(Poster): We prove a new upper bound on the generalization gap of classifiers that are obtained by first using self-supervision to learn a representation r of the training~data, and then fitting a simple (e.g., linear) classifier g to the labels. Specifically, we show that (under the assumptions described below) the generalization gap of such classifiers tends to zero if $\mathsf{C}(g) \ll n$, where $\mathsf{C}(g)$ is an appropriately-defined measure of the simple classifier g's complexity, and n is the number of training samples. We stress that our bound is independent of the complexity of the representation r. We do not make any structural or conditional-independence assumptions on the representation-learning task, which can use the same training dataset that is later used for classification. Rather, we assume that the training procedure satisfies certain natural noise-robustness (adding small amount of label noise causes small degradation in performance) and rationality (getting the wrong label is not better than getting no label at all) conditions that widely hold across many standard architectures. We also conduct an extensive empirical study of the generalization gap and the quantities used in our assumptions for a variety of self-supervision based algorithms, including SimCLR, AMDIM and BigBiGAN, on the CIFAR-10 and ImageNet datasets. We show that, unlike standard supervised classifiers, these algorithms display small generalization gap, and the bounds we prove on this gap are often non vacuous.

Directed Acyclic Graph Neural Networks

- Veronika Thost, Jie Chen
- abstract@open-review(Poster): Graph-structured data ubiquitously appears in science and engineering. Graph neural networks (GNNs) are designed to exploit the relational inductive bias exhibited in graphs; they have been shown to outperform other forms of neural networks in scenarios where structure information supplements node features. The most common GNN architecture aggregates information from neighborhoods based on message passing. Its generality has made it broadly applicable. In this paper, we focus on a special, yet widely used, type of graphs---DAGs---and inject a stronger inductive bias---partial ordering---into the neural network design. We propose the directed acyclic graph neural network, DAGNN, an architecture that processes information according to the flow defined by the partial order. DAGNN can be considered a framework that entails earlier works as special cases (e.g., models for trees and models updating node representations recurrently), but we identify several crucial components that prior architectures lack. We perform comprehensive experiments, including ablation studies, on representative DAG datasets (i.e., source code, neural architectures, and probabilistic graphical models) and demonstrate the superiority of DAGNN over simpler DAG architectures as well as general graph architectures.

The Traveling Observer Model: Multi-task Learning Through Spatial Variable Embeddings

- Elliot Meyerson, Risto Miikkulainen
- abstract@open-review(Spotlight): This paper frames a general prediction system as an observer traveling around a continuous space, measuring values at some locations, and predicting them at others. The observer is completely agnostic about any particular task being solved; it cares only about measurement locations and their values. This perspective leads to a machine learning framework in which seemingly unrelated tasks can be solved by a single model, by embedding their input and output variables into a shared space. An implementation of the framework is developed in which these variable embeddings are learned jointly with internal model parameters. In experiments, the approach is shown to (1) recover intuitive locations of variables in space and time, (2) exploit regularities across related datasets with completely disjoint input and output spaces, and (3) exploit regularities across seemingly unrelated tasks, outperforming task-specific single-task models and multi-task learning alternatives. The results suggest that even seemingly unrelated tasks may originate from similar underlying processes, a fact that the traveling observer model can use to make better predictions.

My Body is a Cage: the Role of Morphology in Graph-Based Incompatible Control

- Vitaly Kurin, Maximilian Igl, Tim Rocktäschel, Wendelin Boehmer, Shimon Whiteson
- abstract@open-review(Poster): Multitask Reinforcement Learning is a promising way to obtain models with better performance, generalisation, data efficiency, and robustness. Most existing work is limited to compatible settings, where the state and action space dimensions are the same across tasks.

Graph Neural Networks (GNN) are one way to address incompatible environments, because they can process graphs of arbitrary size. They also allow practitioners to inject biases encoded in the structure of the input graph. Existing work in graph-based continuous control uses the physical morphology of the agent to construct the input graph, i.e., encoding limb features as node labels and using edges to connect the nodes if their corresponded limbs are physically connected. In this work, we present a series of ablations on existing methods that show that morphological information encoded in the graph does not improve their performance. Motivated by the hypothesis that any benefits GNNs extract from the graph structure are outweighed by difficulties they create for message passing, we also propose Amorpheus, a transformer-based approach. Further results show that, while Amorpheus ignores the morphological information that GNNs encode, it nonetheless substantially outperforms GNN-based methods.

Incremental few-shot learning via vector quantization in deep embedded space

- Kuilin Chen, Chi-Guhn Lee
- abstract@open-review(Poster): The capability of incrementally learning new tasks without forgetting old ones is a challenging problem due to catastrophic forgetting. This challenge becomes greater when novel tasks contain very few labelled training samples. Currently, most methods are dedicated to class-incremental learning and rely on sufficient training data to learn additional weights for newly added classes. Those methods cannot be easily extended to incremental regression tasks and could suffer from severe overfitting when learning few-shot novel tasks. In this study, we propose a nonparametric method in deep embedded space to tackle incremental few-shot learning problems. The knowledge about the learned tasks are compressed into a small number of quantized reference vectors. The proposed method learns new tasks sequentially by adding more reference vectors to the model using few-shot samples in each novel task. For classification problems, we employ the nearest neighbor scheme to make classification on sparsely available data and incorporate intra-class variation, less forgetting regularization and calibration of reference vectors to mitigate catastrophic forgetting. In addition, the proposed learning vector quantization (LVQ) in deep embedded space can be customized as a kernel smoother to handle incremental few-shot regression tasks. Experimental results demonstrate that the proposed method outperforms other state-of-the-art methods in incremental learning.

Revisiting Few-sample BERT Fine-tuning

- Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, Yoav Artzi
- abstract@open-review(Poster): This paper is a study of fine-tuning of BERT contextual representations, with focus on commonly observed instabilities in few-sample scenarios. We identify several factors that cause this instability: the common use of a non-standard optimization method with biased gradient estimation; the limited applicability of significant parts of the BERT network for down-stream tasks; and the prevalent practice of using a pre-determined, and small number of training iterations. We empirically test the impact of these factors, and identify alternative practices that resolve the commonly observed instability of the process. In light of these observations, we re-visit recently proposed methods to improve few-sample fine-tuning with BERT and re-evaluate their effectiveness. Generally, we observe the impact of these methods diminishes significantly with our modified process.

Linear Convergent Decentralized Optimization with Compression

- Xiaorui Liu, Yao Li, Rongrong Wang, Jiliang Tang, Ming Yan
- abstract@open-review(Poster): Communication compression has become a key strategy to speed up distributed optimization. However, existing decentralized algorithms with compression mainly focus on compressing DGD-type algorithms. They are unsatisfactory in terms of convergence rate, stability, and the capability to handle heterogeneous data. Motivated by primal-dual algorithms, this paper proposes the first \underline{L}in\underline{EA}r convergent \underline{D}ecentralized algorithm with compression, LEAD. Our theory describes the coupled dynamics of the inexact primal and dual update as well as compression error, and we provide the first consensus error bound in such settings without assuming bounded gradients. Experiments on convex problems validate our theoretical analysis, and empirical study on deep neural nets shows that LEAD is applicable to non-convex problems.

Learning from Demonstration with Weakly Supervised Disentanglement

- Yordan Hristov, Subramanian Ramamoorthy
- abstract@open-review(Poster): Robotic manipulation tasks, such as wiping with a soft sponge, require control from multiple rich sensory modalities. Human-robot interaction, aimed at teach- ing robots, is difficult in this setting as there is potential for mismatch between human and machine comprehension of the rich data streams. We treat the task of interpretable learning from demonstration as an optimisation problem over a probabilistic generative model. To account for the high-dimensionality of the data, a high-capacity neural network is chosen to represent the model. The latent variables in this model are explicitly aligned with high-level notions and concepts that are manifested in a set of demonstrations. We show that such alignment is best achieved through the use of labels from the end user, in an appropriately restricted vocabulary, in contrast to the conventional approach of the designer picking a prior over the latent variables. Our approach is evaluated in the context of two table-top robot manipulation tasks performed by a PR2 robot – that of dabbing liquids with a sponge (forcefully pressing a sponge and moving it along a surface) and pouring between different containers. The robot provides visual information, arm joint positions and arm joint efforts. We have made videos of the tasks and data available - see supplementary materials at: https://sites.google.com/view/weak-label-lfd.

Incorporating Symmetry into Deep Dynamics Models for Improved Generalization

- Rui Wang, Robin Walters, Rose Yu
- abstract@open-review(Poster): Recent work has shown deep learning can accelerate the prediction of physical dynamics relative to numerical solvers. However, limited physical accuracy and an inability to generalize under distributional shift limit its applicability to the real world. We propose to improve accuracy and generalization by incorporating symmetries into convolutional neural networks. Specifically, we employ a variety of methods each tailored to enforce a different symmetry. Our models are both theoretically and experimentally robust to distributional shift by symmetry group transformations and enjoy favorable sample complexity. We demonstrate the advantage of our approach on a variety of physical dynamics including Rayleigh–Bénard convection and real-world ocean currents and temperatures. Compare with image or text applications, our work is a significant step towards applying equivariant neural networks to high-dimensional systems with complex dynamics.

The Risks of Invariant Risk Minimization

- Elan Rosenfeld, Pradeep Kumar Ravikumar, Andrej Risteski
- abstract@open-review(Poster): Invariant Causal Prediction (Peters et al., 2016) is a technique for out-of-distribution generalization which assumes that some aspects of the data distribution vary across the training set but that the underlying causal mechanisms remain constant. Recently, Arjovsky et al. (2019) proposed Invariant Risk Minimization (IRM), an objective based on this idea for learning deep, invariant features of data which are a complex function of latent variables; many alternatives have subsequently been suggested. However, formal guarantees for all of these works are severely lacking. In this paper, we present the first analysis of classification under the IRM objective—as well as these recently proposed alternatives—under a fairly natural and general model. In the linear case, we show simple conditions under which the optimal solution succeeds or, more often, fails to recover the optimal invariant predictor. We furthermore present the very first results in the non-linear regime: we demonstrate that IRM can fail catastrophically unless the test data is sufficiently similar to the training distribution—this is precisely the issue that it was intended to solve. Thus, in this setting we find that IRM and its alternatives fundamentally do not improve over standard Empirical Risk Minimization.

Scaling Symbolic Methods using Gradients for Neural Model Explanation

- Subham Sekhar Sahoo, Subhashini Venugopalan, Li Li, Rishabh Singh, Patrick Riley
- abstract@open-review(Poster): Symbolic techniques based on Satisfiability Modulo Theory (SMT) solvers have been proposed for analyzing and verifying neural network properties, but their usage has been fairly limited owing to their poor scalability with larger networks. In this work, we propose a technique for combining gradient-based methods with symbolic techniques to scale such analyses and demonstrate its application for model explanation. In particular, we apply this technique to identify minimal regions in an input that are most relevant for a neural network's prediction. Our approach uses gradient information (based on Integrated Gradients) to focus on a subset of neurons in the first layer, which allows our technique to scale to large networks. The corresponding SMT constraints encode the minimal input mask discovery problem such that after masking the input, the activations of the selected neurons are still above a threshold. After solving for the minimal masks, our approach scores the mask regions to generate a relative ordering of the features within the mask. This produces a saliency map which explains" where a model is looking" when making a prediction. We evaluate our technique on three datasets-MNIST, ImageNet, and Beer Reviews, and demonstrate both quantitatively and qualitatively that the regions generated by our approach are sparser and achieve higher saliency scores compared to the gradient-based methods alone. Code and examples are at - https://github.com/google-research/google-research/tree/master/smug_saliency

Contextual Dropout: An Efficient Sample-Dependent Dropout Module

- XINJIE FAN, Shujian Zhang, Korawat Tanwisuth, Xiaoning Qian, Mingyuan Zhou
- abstract@open-review(Poster): Dropout has been demonstrated as a simple and effective module to not only regularize the training process of deep neural networks, but also provide the uncertainty estimation for prediction. However, the quality of uncertainty estimation is highly dependent on the dropout probabilities. Most current models use the same dropout distributions across all data samples due to its simplicity. Despite the potential gains in the flexibility of modeling uncertainty, sample-dependent dropout, on the other hand, is less explored as it often encounters scalability issues or involves non-trivial model changes. In this paper, we propose contextual dropout with an efficient structural design as a simple and scalable sample-dependent dropout module, which can be applied to a wide range of models at the expense of only slightly increased memory and computational cost. We learn the dropout probabilities with a variational objective, compatible with both Bernoulli dropout and Gaussian dropout. We apply the contextual dropout module to various models with applications to image classification and visual question answering and demonstrate the scalability of the method with large-scale datasets, such as ImageNet and VQA 2.0. Our experimental results show that the proposed method outperforms baseline methods in terms of both accuracy and quality of uncertainty estimation.

Contrastive Learning with Hard Negative Samples

- Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, Stefanie Jegelka
- abstract@open-review(Poster): We consider the question: how can you sample good negative examples for contrastive learning? We argue that, as with metric learning, learning contrastive representations benefits from hard negative samples (i.e., points that are difficult to distinguish from an anchor point). The key challenge toward using hard negatives is that contrastive methods must remain unsupervised, making it infeasible to adopt existing negative sampling strategies that use label information. In response, we develop a new class of unsupervised methods for selecting hard negative samples where the user can control the amount of hardness. A limiting case of this sampling results in a representation that tightly clusters each class, and pushes different classes as far apart as possible. The proposed method improves downstream performance across multiple modalities, requires only few additional lines of code to implement, and introduces no computational overhead.

Debiasing Concept-based Explanations with Causal Analysis

- Mohammad Taha Bahadori, David Heckerman
- abstract@open-review(Poster): Concept-based explanation approach is a popular model interpertability tool because it expresses the reasons for a model's predictions in terms of concepts that are meaningful for the domain experts. In this work, we study the problem of the concepts being correlated with confounding information in the features. We propose a new causal prior graph for modeling the impacts of unobserved variables and a method to remove the impact of confounding information and noise using a two-stage regression technique borrowed from the instrumental variable literature. We also model the completeness of the concepts set and show that our debiasing method works when the concepts are not complete. Our synthetic and real-world experiments demonstrate the success of our method in removing biases and improving the ranking of the concepts in terms of their contribution to the explanation of the predictions.

Self-training For Few-shot Transfer Across Extreme Task Differences

- Cheng Perng Phoo, Bharath Hariharan
- abstract@open-review(Oral): Most few-shot learning techniques are pre-trained on a large, labeled “base dataset”. In problem domains where such large labeled datasets are not available for pre-training (e.g., X-ray, satellite images), one must resort to pre-training in a different “source” problem domain (e.g., ImageNet), which can be very different from the desired target task. Traditional few-shot and transfer learning techniques fail in the presence of such extreme differences between the source and target tasks. In this paper, we present a simple and effective solution to tackle this extreme domain gap: self-training a source domain representation on unlabeled data from the target domain. We show that this improves one-shot performance on the target domain by 2.9 points on average on the challenging BSCD-FSL benchmark consisting of datasets from multiple domains.

Risk-Averse Offline Reinforcement Learning

- Núria Armengol Urpí, Sebastian Curi, Andreas Krause
- abstract@open-review(Poster): Training Reinforcement Learning (RL) agents in high-stakes applications might be too prohibitive due to the risk associated to exploration. Thus, the agent can only use data previously collected by safe policies. While previous work considers optimizing the average performance using offline data, we focus on optimizing a risk-averse criteria, namely the CVaR. In particular, we present the Offline Risk-Averse Actor-Critic (O-RAAC), a model-free RL algorithm that is able to learn risk-averse policies in a fully offline setting. We show that O-RAAC learns policies with higher CVaR than risk-neutral approaches in different robot control tasks. Furthermore, considering risk-averse criteria guarantees distributional robustness of the average performance with respect to particular distribution shifts. We demonstrate empirically that in the presence of natural distribution-shifts, O-RAAC learns policies with good average performance.

Parameter Efficient Multimodal Transformers for Video Representation Learning

- Sangho Lee, Youngjae Yu, Gunhee Kim, Thomas Breuel, Jan Kautz, Yale Song
- abstract@open-review(Poster): The recent success of Transformers in the language domain has motivated adapting it to a multimodal setting, where a new visual model is trained in tandem with an already pretrained language model. However, due to the excessive memory requirements from Transformers, existing work typically fixes the language model and train only the vision module, which limits its ability to learn cross-modal information in an end-to-end manner. In this work, we focus on reducing the parameters of multimodal Transformers in the context of audio-visual video representation learning. We alleviate the high memory requirement by sharing the parameters of Transformers across layers and modalities; we decompose the Transformer into modality-specific and modality-shared parts so that the model learns the dynamics of each modality both individually and together, and propose a novel

parameter sharing scheme based on low-rank approximation. We show that our approach reduces parameters of the Transformers up to 97%, allowing us to train our model end-to-end from scratch. We also propose a negative sampling approach based on an instance similarity measured on the CNN embedding space that our model learns together with the Transformers. To demonstrate our approach, we pretrain our model on 30-second clips (480 frames) from Kinetics-700 and transfer it to audio-visual classification tasks.

Tradeoffs in Data Augmentation: An Empirical Study

- Raphael Gontijo-Lopes, Sylvia Smullin, Ekin Dogus Cubuk, Ethan Dyer
- abstract@open-review(Poster): Though data augmentation has become a standard component of deep neural network training, the underlying mechanism behind the effectiveness of these techniques remains poorly understood. In practice, augmentation policies are often chosen using heuristics of distribution shift or augmentation diversity. Inspired by these, we conduct an empirical study to quantify how data augmentation improves model generalization. We introduce two interpretable and easy-to-compute measures: Affinity and Diversity. We find that augmentation performance is predicted not by either of these alone but by jointly optimizing the two.

Concept Learners for Few-Shot Learning

- Kaidi Cao, Maria Brbic, Jure Leskovec
- abstract@open-review(Poster): Developing algorithms that are able to generalize to a novel task given only a few labeled examples represents a fundamental challenge in closing the gap between machine- and human-level performance. The core of human cognition lies in the structured, reusable concepts that help us to rapidly adapt to new tasks and provide reasoning behind our decisions. However, existing meta-learning methods learn complex representations across prior labeled tasks without imposing any structure on the learned representations. Here we propose COMET, a meta-learning method that improves generalization ability by learning to learn along human-interpretable concept dimensions. Instead of learning a joint unstructured metric space, COMET learns mappings of high-level concepts into semi-structured metric spaces, and effectively combines the outputs of independent concept learners. We evaluate our model on few-shot tasks from diverse domains, including fine-grained image classification, document categorization and cell type annotation on a novel dataset from a biological domain developed in our work. COMET significantly outperforms strong meta-learning baselines, achieving 6-15% relative improvement on the most challenging 1-shot learning tasks, while unlike existing methods providing interpretations behind the model's predictions.

Neural Mechanics: Symmetry and Broken Conservation Laws in Deep Learning Dynamics

- Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel LK Yamins, Hidenori Tanaka
- abstract@open-review(Poster): Understanding the dynamics of neural network parameters during training is one of the key challenges in building a theoretical foundation for deep learning. A central obstacle is that the motion of a network in high-dimensional parameter space undergoes discrete finite steps along complex stochastic gradients derived from real-world datasets. We circumvent this obstacle through a unifying theoretical framework based on intrinsic symmetries embedded in a network's architecture that are present for any dataset. We show that any such symmetry imposes stringent geometric constraints on gradients and Hessians, leading to an associated conservation law in the continuous-time limit of stochastic gradient descent (SGD), akin to Noether's theorem in physics. We further show that finite learning rates used in practice can actually break these symmetry induced conservation laws. We apply tools from finite difference methods to derive modified gradient flow, a differential equation that better approximates the numerical trajectory taken by SGD at finite learning rates. We combine modified gradient flow with our framework of symmetries to derive exact integral expressions for the dynamics of certain parameter combinations. We empirically validate our analytic expressions for learning dynamics on VGG-16 trained on Tiny ImageNet. Overall, by exploiting symmetry, our work demonstrates that we can analytically describe the learning dynamics of various parameter combinations at finite learning rates and batch sizes for state of the art architectures trained on any dataset.

Federated Learning Based on Dynamic Regularization

- Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, Venkatesh Saligrama
- abstract@open-review(Oral): We propose a novel federated learning method for distributively training neural network models, where the server orchestrates cooperation between a subset of randomly chosen devices in each round. We view Federated Learning problem primarily from a communication perspective and allow more device level computations to save transmission costs. We point out a fundamental dilemma, in that the minima of the local-device level empirical loss are inconsistent with those of the global empirical loss. Different from recent prior works, that either attempt inexact minimization or utilize devices for parallelizing gradient computation, we propose a dynamic regularizer for each device at each round, so that in the limit the global and device solutions are aligned. We demonstrate both through empirical results on real and synthetic data as well as analytical results that our scheme leads to efficient training, in both convex and non-convex settings, while being fully agnostic to device heterogeneity and robust to large number of devices, partial participation and unbalanced data.

Fair Mixup: Fairness via Interpolation

- Ching-Yao Chuang, Youssef Mroueh
- abstract@open-review(Poster): Training classifiers under fairness constraints such as group fairness, regularizes the disparities of predictions between the groups. Nevertheless, even though the constraints are satisfied during training, they might not generalize at evaluation time. To improve the generalizability of fair classifiers, we propose fair mixup, a new data augmentation strategy for imposing the fairness constraint. In particular, we show that fairness can be achieved by regularizing the models on paths of interpolated samples between the groups. We use mixup, a powerful data augmentation strategy to generate these interpolates. We analyze fair mixup and empirically show that it ensures a better generalization for both accuracy and fairness measurement in tabular, vision, and language benchmarks.

Fidelity-based Deep Adiabatic Scheduling

- Eli Ovits, Lior Wolf
- abstract@open-review(Spotlight): Adiabatic quantum computation is a form of computation that acts by slowly interpolating a quantum system between an easy to prepare initial state and a final state that represents a solution to a given computational problem. The choice of the interpolation schedule is critical to the performance: if at a certain time point, the evolution is too rapid, the system has a high probability to transfer to a higher energy state, which does not represent a solution to the problem. On the other hand, an evolution that is too slow leads to a loss of computation time and increases the probability of failure due to decoherence. In this work, we train deep neural models to produce optimal schedules that are conditioned on the problem at hand. We consider two types of problem representation: the Hamiltonian form, and the Quadratic Unconstrained Binary Optimization (QUBO) form. A novel loss function that scores schedules according to their approximated success probability is introduced. We benchmark our approach on random QUBO problems, Grover search, 3-SAT, and MAX-CUT problems and show that our approach outperforms, by a sizable margin, the linear schedules as well as alternative approaches that were very recently proposed.

Heating up decision boundaries: isocapacitory saturation, adversarial scenarios and generalization bounds

- Bogdan Georgiev, Lukas Franken, Mayukh Mukherjee

- abstract@open-review(Poster): In the present work we study classifiers' decision boundaries via Brownian motion processes in ambient data space and associated probabilistic techniques. Intuitively, our ideas correspond to placing a heat source at the decision boundary and observing how effectively the sample points warm up. We are largely motivated by the search for a soft measure that sheds further light on the decision boundary's geometry. En route, we bridge aspects of potential theory and geometric analysis (Maz'ya 2011, Grigor'Yan and Saloff-Coste 2002) with active fields of ML research such as adversarial examples and generalization bounds. First, we focus on the geometric behavior of decision boundaries in the light of adversarial attack/defense mechanisms. Experimentally, we observe a certain capacitory trend over different adversarial defense strategies: decision boundaries locally become flatter as measured by isoperimetric inequalities (Ford et al 2019); however, our more sensitive heat-diffusion metrics extend this analysis and further reveal that some non-trivial geometry invisible to plain distance-based methods is still preserved. Intuitively, we provide evidence that the decision boundaries nevertheless retain many persistent "wiggly and fuzzy" regions on a finer scale. Second, we show how Brownian hitting probabilities translate to soft generalization bounds which are in turn connected to compression and noise stability (Arora et al 2018), and these bounds are significantly stronger if the decision boundary has controlled geometric features.

Deciphering and Optimizing Multi-Task Learning: a Random Matrix Approach

- Malik Tiomoko, Hafiz Tiomoko Ali, Romain Couillet
- abstract@open-review(Spotlight): This article provides theoretical insights into the inner workings of multi-task and transfer learning methods, by studying the tractable least-square support vector machine multi-task learning (LS-SVM MTL) method, in the limit of large (p) and numerous (n) data. By a random matrix analysis applied to a Gaussian mixture data model, the performance of MTL LS-SVM is shown to converge, as $n,p\to\infty$, to a deterministic limit involving simple (small-dimensional) statistics of the data.

We prove (i) that the standard MTL LS-SVM algorithm is in general strongly biased and may dramatically fail (to the point that individual single-task LS-SVMs may outperform the MTL approach, even for quite resembling tasks): our analysis provides a simple method to correct these biases, and that we reveal (ii) the sufficient statistics at play in the method, which can be efficiently estimated, even for quite small datasets. The latter result is exploited to automatically optimize the hyperparameters without resorting to any cross-validation procedure.

Experiments on popular datasets demonstrate that our improved MTL LS-SVM method is computationally-efficient and outperforms sometimes much more elaborate state-of-the-art multi-task and transfer learning techniques.

Influence Functions in Deep Learning Are Fragile

- Samyadeep Basu, Phil Pope, Soheil Feizi
- abstract@open-review(Poster): Influence functions approximate the effect of training samples in test-time predictions and have a wide variety of applications in machine learning interpretability and uncertainty estimation. A commonly-used (first-order) influence function can be implemented efficiently as a post-hoc method requiring access only to the gradients and Hessian of the model. For linear models, influence functions are well-defined due to the convexity of the underlying loss function and are generally accurate even across difficult settings where model changes are fairly large such as estimating group influences. Influence functions, however, are not well-understood in the context of deep learning with non-convex loss functions. In this paper, we provide a comprehensive and large-scale empirical study of successes and failures of influence functions in neural network models trained on datasets such as Iris, MNIST, CIFAR-10 and ImageNet. Through our extensive experiments, we show that the network architecture, its depth and width, as well as the extent of model parameterization and regularization techniques have strong effects in the accuracy of influence functions. In particular, we find that (i) influence estimates are fairly accurate for shallow networks, while for deeper networks the estimates are often erroneous; (ii) for certain network architectures and datasets, training with weight-decay regularization is important to get high-quality influence estimates; and (iii) the accuracy of influence estimates can vary significantly depending on the examined test points. These results suggest that in general influence functions in deep learning are fragile and call for developing improved influence estimation methods to mitigate these issues in non-convex setups.

Few-Shot Learning via Learning the Representation, Provably

- Simon Shaolei Du, Wei Hu, Sham M. Kakade, Jason D. Lee, Qi Lei
- abstract@open-review(Poster): This paper studies few-shot learning via representation learning, where one uses T source tasks with n_1 data per task to learn a representation in order to reduce the sample complexity of a target task for which there is only $n_2 (\ll n_1)$ data. Specifically, we focus on the setting where there exists a good common representation between source and target, and our goal is to understand how much a sample size reduction is possible. First, we study the setting where this common representation is low-dimensional and provide a risk bound of $\tilde{O}(\frac{dk}{n_1T} + \frac{k}{n_2})$ on the target task for the linear representation class; here d is the ambient input dimension and $k (\ll d)$ is the dimension of the representation. This result bypasses the $\Omega(\frac{1}{T})$ barrier under the i.i.d. task assumption, and can capture the desired property that all n_1T samples from source tasks can be \emph{pooled} together for representation learning. We further extend this result to handle a general representation function class and obtain a similar result. Next, we consider the setting where the common representation may be high-dimensional but is capacity-constrained (say in norm); here, we again demonstrate the advantage of representation learning in both high-dimensional linear regression and neural networks, and show that representation learning can fully utilize all n_1T samples from source tasks.

Self-Supervised Learning of Compressed Video Representations

- Youngjae Yu, Sangho Lee, Gunhee Kim, Yale Song
- abstract@open-review(Poster): Self-supervised learning of video representations has received great attention. Existing methods typically require frames to be decoded before being processed, which increases compute and storage requirements and ultimately hinders large-scale training. In this work, we propose an efficient self-supervised approach to learn video representations by eliminating the expensive decoding step. We use a three-stream video architecture that encodes I-frames and P-frames of a compressed video. Unlike existing approaches that encode I-frames and P-frames individually, we propose to jointly encode them by establishing bidirectional dynamic connections across streams. To enable self-supervised learning, we propose two pretext tasks that leverage the multimodal nature (RGB, motion vector, residuals) and the internal GOP structure of compressed videos. The first task asks our network to predict zeroth-order motion statistics in a spatio-temporal pyramid; the second task asks correspondence types between I-frames and P-frames after applying temporal transformations. We show that our approach achieves competitive performance on compressed video recognition both in supervised and self-supervised regimes.

Flowtron: an Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis

- Rafael Valle, Kevin J. Shih, Ryan Prenger, Bryan Catanzaro
- abstract@open-review(Poster): In this paper we propose Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis with style transfer and speech variation. Flowtron borrows insights from Autoregressive Flows and revamps Tacotron 2 in order to provide high-quality and expressive mel-spectrogram synthesis. Flowtron is optimized by maximizing the likelihood of the training data, which makes training simple and stable. Flowtron learns an invertible mapping of data to a latent space that can be used to modulate many aspects of speech synthesis (timbre, expressivity, accent). Our mean opinion scores (MOS) show that Flowtron matches state-of-the-art TTS models in terms of speech quality. We provide results on speech variation, interpolation over time between samples and style transfer between seen and unseen speakers. Code and pre-trained models are publicly available at \href{https://github.com/NVIDIA/flowtron}{https://github.com/NVIDIA/flowtron}.

R-GAP: Recursive Gradient Attack on Privacy

- Junyi Zhu, Matthew B. Blaschko
- abstract@open-review(Poster): Federated learning frameworks have been regarded as a promising approach to break the dilemma between demands on privacy and the promise of learning from large collections of distributed data. Many such frameworks only ask collaborators to share their local update of a common model, i.e. gradients with respect to locally stored data, instead of exposing their raw data to other collaborators. However, recent optimization-based gradient attacks show that raw data can often be accurately recovered from gradients. It has been shown that minimizing the Euclidean distance between true gradients and those calculated from estimated data is often effective in fully recovering private data. However, there is a fundamental lack of theoretical understanding of how and when gradients can lead to unique recovery of original data. Our research fills this gap by providing a closed-form recursive procedure to recover data from gradients in deep neural networks. We name it Recursive Gradient Attack on Privacy (R-GAP). Experimental results demonstrate that R-GAP works as well as or even better than optimization-based approaches at a fraction of the computation under certain conditions. Additionally, we propose a Rank Analysis method, which can be used to estimate the risk of gradient attacks inherent in certain network architectures, regardless of whether an optimization-based or closed-form-recursive attack is used. Experimental results demonstrate the utility of the rank analysis towards improving the network's security. Source code is available for download from https://github.com/JunyiZhu-AI/R-GAP.

Why resampling outperforms reweighting for correcting sampling bias with stochastic gradients

- Jing An, Lexing Ying, Yuhua Zhu
- abstract@open-review(Poster): A data set sampled from a certain population is biased if the subgroups of the population are sampled at proportions that are significantly different from their underlying proportions. Training machine learning models on biased data sets requires correction techniques to compensate for the bias. We consider two commonly-used techniques, resampling and reweighting, that rebalance the proportions of the subgroups to maintain the desired objective function. Though statistically equivalent, it has been observed that resampling outperforms reweighting when combined with stochastic gradient algorithms. By analyzing illustrative examples, we explain the reason behind this phenomenon using tools from dynamical stability and stochastic asymptotics. We also present experiments from regression, classification, and off-policy prediction to demonstrate that this is a general phenomenon. We argue that it is imperative to consider the objective function design and the optimization algorithm together while addressing the sampling bias.

Learning with Instance-Dependent Label Noise: A Sample Sieve Approach

- Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, Yang Liu
- abstract@open-review(Poster): Human-annotated labels are often prone to noise, and the presence of such noise will degrade the performance of the resulting deep neural network (DNN) models. Much of the literature (with several recent exceptions) of learning with noisy labels focuses on the case when the label noise is independent of features. Practically, annotations errors tend to be instance-dependent and often depend on the difficulty levels of recognizing a certain task. Applying existing results from instance-independent settings would require a significant amount of estimation of noise rates. Therefore, providing theoretically rigorous solutions for learning with instance-dependent label noise remains a challenge. In this paper, we propose CORES^{2} (COnfidence REgularized Sample Sieve), which progressively sieves out corrupted examples. The implementation of CORES^{2} does not require specifying noise rates and yet we are able to provide theoretical guarantees of CORES^{2} in filtering out the corrupted examples. This high-quality sample sieve allows us to treat clean examples and the corrupted ones separately in training a DNN solution, and such a separation is shown to be advantageous in the instance-dependent noise setting. We demonstrate the performance of CORES^{2} on CIFAR10 and CIFAR100 datasets with synthetic instance-dependent label noise and Clothing1M with real-world human noise. As of independent interests, our sample sieve provides a generic machinery for anatomizing noisy datasets and provides a flexible interface for various robust training techniques to further improve the performance. Code is available at https://github.com/UCSC-REAL/cores.

Unsupervised Audiovisual Synthesis via Exemplar Autoencoders

- Kangle Deng, Aayush Bansal, Deva Ramanan
- abstract@open-review(Poster): We present an unsupervised approach that converts the input speech of any individual into audiovisual streams of potentially-infinitely many output speakers. Our approach builds on simple autoencoders that project out-of-sample data onto the distribution of the training set. We use exemplar autoencoders to learn the voice, stylistic prosody, and visual appearance of a specific target exemplar speech. In contrast to existing methods, the proposed approach can be easily extended to an arbitrarily large number of speakers and styles using only 3 minutes of target audio-video data, without requiring any training data for the input speaker. To do so, we learn audiovisual bottleneck representations that capture the structured linguistic content of speech. We outperform prior approaches on both audio and video synthesis.

Single-Timescale Actor-Critic Provably Finds Globally Optimal Policy

- Zuyue Fu, Zhuoran Yang, Zhaoran Wang
- abstract@open-review(Poster): We study the global convergence and global optimality of actor-critic, one of the most popular families of reinforcement learning algorithms. While most existing works on actor-critic employ bi-level or two-timescale updates, we focus on the more practical single-timescale setting, where the actor and critic are updated simultaneously. Specifically, in each iteration, the critic update is obtained by applying the Bellman evaluation operator only once while the actor is updated in the policy gradient direction computed using the critic. Moreover, we consider two function approximation settings where both the actor and critic are represented by linear or deep neural networks. For both cases, we prove that the actor sequence converges to a globally optimal policy at a sublinear $O(K^{-1/2})$ rate, where K is the number of iterations. To the best of our knowledge, we establish the rate of convergence and global optimality of single-timescale actor-critic with linear function approximation for the first time. Moreover, under the broader scope of policy optimization with nonlinear function approximation, we prove that actor-critic with deep neural network finds the globally optimal policy at a sublinear rate for the first time.

Wandering within a world: Online contextualized few-shot learning

- Mengye Ren, Michael Louis Iuzzolino, Michael Curtis Mozer, Richard Zemel
- abstract@open-review(Poster): We aim to bridge the gap between typical human and machine-learning environments by extending the standard framework of few-shot learning to an online, continual setting. In this setting, episodes do not have separate training and testing phases, and instead models are evaluated online while learning novel classes. As in the real world, where the presence of spatiotemporal context helps us retrieve learned skills in the past, our online few-shot learning setting also features an underlying context that changes throughout time. Object classes are correlated within a context and inferring the correct context can lead to better performance. Building upon this setting, we propose a new few-shot learning dataset based on large scale indoor imagery that mimics the visual experience of an agent wandering within a world. Furthermore, we convert popular few-shot learning approaches into online versions and we also propose a new model that can make use of spatiotemporal contextual information from the recent past.

Learning-based Support Estimation in Sublinear Time

- Talya Eden, Piotr Indyk, Shyam Narayanan, Ronitt Rubinfeld, Sandeep Silwal, Tal Wagner

- abstract@open-review(Spotlight): We consider the problem of estimating the number of distinct elements in a large data set (or, equivalently, the support size of the distribution induced by the data set) from a random sample of its elements. The problem occurs in many applications, including biology, genomics, computer systems and linguistics. A line of research spanning the last decade resulted in algorithms that estimate the support up to $ \pm \varepsilon n$ from a sample of size $O(\log^2(1/\varepsilon) \cdot n/\log n)$, where n is the data set size. Unfortunately, this bound is known to be tight, limiting further improvements to the complexity of this problem. In this paper we consider estimation algorithms augmented with a machine-learning-based predictor that, given any element, returns an estimation of its frequency. We show that if the predictor is correct up to a constant approximation factor, then the sample complexity can be reduced significantly, to $$ \ \log (1/\varepsilon) \cdot n^{1-\Theta(1/\log(1/\varepsilon))}. $$ We evaluate the proposed algorithms on a collection of data sets, using the neural-network based estimators from {Hsu et al, ICLR'19} as predictors. Our experiments demonstrate substantial (up to 3x) improvements in the estimation accuracy compared to the state of the art algorithm.

Neural Delay Differential Equations

- Qunxi Zhu, Yao Guo, Wei Lin
- abstract@open-review(Poster): Neural Ordinary Differential Equations (NODEs), a framework of continuous-depth neural networks, have been widely applied, showing exceptional efficacy in coping with some representative datasets. Recently, an augmented framework has been successfully developed for conquering some limitations emergent in application of the original framework. Here we propose a new class of continuous-depth neural networks with delay, named as Neural Delay Differential Equations (NDDEs), and, for computing the corresponding gradients, we use the adjoint sensitivity method to obtain the delayed dynamics of the adjoint. Since the differential equations with delays are usually seen as dynamical systems of infinite dimension possessing more fruitful dynamics, the NDDEs, compared to the NODEs, own a stronger capacity of nonlinear representations. Indeed, we analytically validate that the NDDEs are of universal approximators, and further articulate an extension of the NDDEs, where the initial function of the NDDEs is supposed to satisfy ODEs. More importantly, we use several illustrative examples to demonstrate the outstanding capacities of the NDDEs and the NDDEs with ODEs' initial value. More precisely, (1) we successfully model the delayed dynamics where the trajectories in the lower-dimensional phase space could be mutually intersected, while the traditional NODEs without any argumentation are not directly applicable for such modeling, and (2) we achieve lower loss and higher accuracy not only for the data produced synthetically by complex models but also for the real-world image datasets, i.e., CIFAR10, MNIST and SVHN. Our results on the NDDEs reveal that appropriately articulating the elements of dynamical systems into the network design is truly beneficial to promoting the network performance.

Dance Revolution: Long-Term Dance Generation with Music via Curriculum Learning

- Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, Daxin Jiang
- abstract@open-review(Poster): Dancing to music is one of human's innate abilities since ancient times. In machine learning research, however, synthesizing dance movements from music is a challenging problem. Recently, researchers synthesize human motion sequences through autoregressive models like recurrent neural network (RNN). Such an approach often generates short sequences due to an accumulation of prediction errors that are fed back into the neural network. This problem becomes even more severe in the long motion sequence generation. Besides, the consistency between dance and music in terms of style, rhythm and beat is yet to be taken into account during modeling. In this paper, we formalize the music-driven dance generation as a sequence-to-sequence learning problem and devise a novel seq2seq architecture to efficiently process long sequences of music features and capture the fine-grained correspondence between music and dance. Furthermore, we propose a novel curriculum learning strategy to alleviate error accumulation of autoregressive models in long motion sequence generation, which gently changes the training process from a fully guided teacher-forcing scheme using the previous ground-truth movements, towards a less guided autoregressive scheme mostly using the generated movements instead. Extensive experiments show that our approach significantly outperforms the existing state-of-the-arts on automatic metrics and human evaluation. We also make a demo video to demonstrate the superior performance of our proposed approach at https://www.youtube.com/watch?v=lmE20MEheZ8.

Learning Parametrised Graph Shift Operators

- George Dasoulas, Johannes F. Lutzeyer, Michalis Vazirgiannis
- abstract@open-review(Poster): In many domains data is currently represented as graphs and therefore, the graph representation of this data becomes increasingly important in machine learning. Network data is, implicitly or explicitly, always represented using a graph shift operator (GSO) with the most common choices being the adjacency, Laplacian matrices and their normalisations. In this paper, a novel parametrised GSO (PGSO) is proposed, where specific parameter values result in the most commonly used GSOs and message-passing operators in graph neural network (GNN) frameworks. The PGSO is suggested as a replacement of the standard GSOs that are used in state-of-the-art GNN architectures and the optimisation of the PGSO parameters is seamlessly included in the model training. It is proved that the PGSO has real eigenvalues and a set of real eigenvectors independent of the parameter values and spectral bounds on the PGSO are derived. PGSO parameters are shown to adapt to the sparsity of the graph structure in a study on stochastic blockmodel networks, where they are found to automatically replicate the GSO regularisation found in the literature. On several real-world datasets the accuracy of state-of-the-art GNN architectures is improved by the inclusion of the PGSO in both node- and graph-classification tasks.

Unlearnable Examples: Making Personal Data Unexploitable

- Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, Yisen Wang
- abstract@open-review(Spotlight): The volume of "free" data on the internet has been key to the current success of deep learning. However, it also raises privacy concerns about the unauthorized exploitation of personal data for training commercial models. It is thus crucial to develop methods to prevent unauthorized data exploitation. This paper raises the question: can data be made unlearnable for deep learning models? We present a type of error-minimizing noise that can indeed make training examples unlearnable. Error-minimizing noise is intentionally generated to reduce the error of one or more of the training example(s) close to zero, which can trick the model into believing there is "nothing" to learn from these example(s). The noise is restricted to be imperceptible to human eyes, and thus does not affect normal data utility. We empirically verify the effectiveness of error-minimizing noise in both sample-wise and class-wise forms. We also demonstrate its flexibility under extensive experimental settings and practicability in a case study of face recognition. Our work establishes an important first step towards making personal data unexploitable to deep learning models.

Rao-Blackwellizing the Straight-Through Gumbel-Softmax Gradient Estimator

- Max B Paulus, Chris J. Maddison, Andreas Krause
- abstract@open-review(Oral): Gradient estimation in models with discrete latent variables is a challenging problem, because the simplest unbiased estimators tend to have high variance. To counteract this, modern estimators either introduce bias, rely on multiple function evaluations, or use learned, input-dependent baselines. Thus, there is a need for estimators that require minimal tuning, are computationally cheap, and have low mean squared error. In this paper, we show that the variance of the straight-through variant of the popular Gumbel-Softmax estimator can be reduced through Rao-Blackwellization without increasing the number of function evaluations. This provably reduces the mean squared error. We empirically demonstrate that this leads to variance reduction, faster convergence, and generally improved performance in two unsupervised latent variable models.

DOP: Off-Policy Multi-Agent Decomposed Policy Gradients

- Yihan Wang, Beining Han, Tonghan Wang, Heng Dong, Chongjie Zhang
- abstract@open-review(Poster): Multi-agent policy gradient (MAPG) methods recently witness vigorous progress. However, there is a significant performance discrepancy between MAPG methods and state-of-the-art multi-agent value-based approaches. In this paper, we investigate causes that

hinder the performance of MAPG algorithms and present a multi-agent decomposed policy gradient method (DOP). This method introduces the idea of value function decomposition into the multi-agent actor-critic framework. Based on this idea, DOP supports efficient off-policy learning and addresses the issue of centralized-decentralized mismatch and credit assignment in both discrete and continuous action spaces. We formally show that DOP critics have sufficient representational capability to guarantee convergence. In addition, empirical evaluations on the StarCraft II micromanagement benchmark and multi-agent particle environments demonstrate that DOP outperforms both state-of-the-art value-based and policy-based multi-agent reinforcement learning algorithms. Demonstrative videos are available at https://sites.google.com/view/dop-mapg/.

Unsupervised Discovery of 3D Physical Objects from Video

- Yilun Du, Kevin A. Smith, Tomer Ullman, Joshua B. Tenenbaum, Jiajun Wu
- abstract@open-review(Poster): We study the problem of unsupervised physical object discovery. While existing frameworks aim to decompose scenes into 2D segments based off each object's appearance, we explore how physics, especially object interactions, facilitates disentangling of 3D geometry and position of objects from video, in an unsupervised manner. Drawing inspiration from developmental psychology, our Physical Object Discovery Network (POD-Net) uses both multi-scale pixel cues and physical motion cues to accurately segment observable and partially occluded objects of varying sizes, and infer properties of those objects. Our model reliably segments objects on both synthetic and real scenes. The discovered object properties can also be used to reason about physical events.

Exploring Balanced Feature Spaces for Representation Learning

- Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, Jiashi Feng
- abstract@open-review(Poster): Existing self-supervised learning (SSL) methods are mostly applied for training representation models from artificially balanced datasets (e.g., ImageNet). It is unclear how well they will perform in the practical scenarios where datasets are often imbalanced w.r.t. the classes. Motivated by this question, we conduct a series of studies on the performance of self-supervised contrastive learning and supervised learning methods over multiple datasets where training instance distributions vary from a balanced one to a long-tailed one. Our findings are quite intriguing. Different from supervised methods with large performance drop, the self-supervised contrastive learning methods perform stably well even when the datasets are heavily imbalanced. This motivates us to explore the balanced feature spaces learned by contrastive learning, where the feature representations present similar linear separability w.r.t. all the classes. Our further experiments reveal that a representation model generating a balanced feature space can generalize better than that yielding an imbalanced one across multiple settings. Inspired by these insights, we develop a novel representation learning method, called k-positive contrastive learning. It effectively combines strengths of the supervised method and the contrastive learning method to learn representations that are both discriminative and balanced. Extensive experiments demonstrate its superiority on multiple recognition tasks. Remarkably, it achieves new state-of-the-art on challenging long-tailed recognition benchmarks. Code and models will be released.

Auxiliary Learning by Implicit Differentiation

- Aviv Navon, Idan Achituve, Haggai Maron, Gal Chechik, Ethan Fetaya
- abstract@open-review(Poster): Training neural networks with auxiliary tasks is a common practice for improving the performance on a main task of interest. Two main challenges arise in this multi-task learning setting: (i) designing useful auxiliary tasks; and (ii) combining auxiliary tasks into a single coherent loss. Here, we propose a novel framework, AuxiLearn, that targets both challenges based on implicit differentiation. First, when useful auxiliaries are known, we propose learning a network that combines all losses into a single coherent objective function. This network can learn non-linear interactions between tasks. Second, when no useful auxiliary task is known, we describe how to learn a network that generates a meaningful, novel auxiliary task. We evaluate AuxiLearn in a series of tasks and domains, including image segmentation and learning with attributes in the low data regime, and find that it consistently outperforms competing methods.

On the Bottleneck of Graph Neural Networks and its Practical Implications

- Uri Alon, Eran Yahav
- abstract@open-review(Poster): Since the proposal of the graph neural network (GNN) by Gori et al. (2005) and Scarselli et al. (2008), one of the major problems in training GNNs was their struggle to propagate information between distant nodes in the graph. We propose a new explanation for this problem: GNNs are susceptible to a bottleneck when aggregating messages across a long path. This bottleneck causes the over-squashing of exponentially growing information into fixed-size vectors. As a result, GNNs fail to propagate messages originating from distant nodes and perform poorly when the prediction task depends on long-range interaction. In this paper, we highlight the inherent problem of over-squashing in GNNs: we demonstrate that the bottleneck hinders popular GNNs from fitting long-range signals in the training data; we further show that GNNs that absorb incoming edges equally, such as GCN and GIN, are more susceptible to over-squashing than GAT and GGNN; finally, we show that prior work, which extensively tuned GNN models of long-range problems, suffers from over-squashing, and that breaking the bottleneck improves their state-of-the-art results without any tuning or additional weights. Our code is available at https://github.com/tech-srl/bottleneck/ .

The Role of Momentum Parameters in the Optimal Convergence of Adaptive Polyak's Heavy-ball Methods

- Wei Tao, Sheng Long, Gaowei Wu, Qing Tao
- abstract@open-review(Poster): The adaptive stochastic gradient descent (SGD) with momentum has been widely adopted in deep learning as well as convex optimization. In practice, the last iterate is commonly used as the final solution. However, the available regret analysis and the setting of constant momentum parameters only guarantee the optimal convergence of the averaged solution. In this paper, we fill this theory-practice gap by investigating the convergence of the last iterate (referred to as {\it individual convergence}), which is a more difficult task than convergence analysis of the averaged solution. Specifically, in the constrained convex cases, we prove that the adaptive Polyak's Heavy-ball (HB) method, in which the step size is only updated using the exponential moving average strategy, attains an individual convergence rate of $O(\frac{1}{\sqrt{t}})$, as opposed to that of $O(\frac{\log t}{\sqrt {t}})$ of SGD, where t is the number of iterations. Our new analysis not only shows how the HB momentum and its time-varying weight help us to achieve the acceleration in convex optimization but also gives valuable hints how the momentum parameters should be scheduled in deep learning. Empirical results validate the correctness of our convergence analysis in optimizing convex functions and demonstrate the improved performance of the adaptive HB methods in training deep networks.

How Benign is Benign Overfitting ?

- Amartya Sanyal, Puneet K. Dokania, Varun Kanade, Philip Torr
- abstract@open-review(Spotlight): We investigate two causes for adversarial vulnerability in deep neural networks: bad data and (poorly) trained models. When trained with SGD, deep neural networks essentially achieve zero training error, even in the presence of label noise, while also exhibiting good generalization on natural test data, something referred to as benign overfitting (Bartlett et al., 2020; Chatterji & Long, 2020). However, these models are vulnerable to adversarial attacks. We identify label noise as one of the causes for adversarial vulnerability, and provide theoretical and empirical evidence in support of this. Surprisingly, we find several instances of label noise in datasets such as MNIST and CIFAR, and that robustly trained models incur training error on some of these, i.e. they don’t fit the noise. However, removing noisy labels alone does not suffice to achieve adversarial robustness. We conjecture that in part sub-optimal representation learning is also responsible for adversarial vulnerability. By means of simple theoretical setups, we show how the choice of representation can drastically affect adversarial robustness.

Entropic gradient descent algorithms and wide flat minima

- Fabrizio Pittorino, Carlo Lucibello, Christoph Feinauer, Gabriele Perugini, Carlo Baldassi, Elizaveta Demyanenko, Riccardo Zecchina
- abstract@open-review(Poster): The properties of flat minima in the empirical risk landscape of neural networks have been debated for some time. Increasing evidence suggests they possess better generalization capabilities with respect to sharp ones. In this work we first discuss the relationship between alternative measures of flatness: The local entropy, which is useful for analysis and algorithm development, and the local energy, which is easier to compute and was shown empirically in extensive tests on state-of-the-art networks to be the best predictor of generalization capabilities. We show semi-analytically in simple controlled scenarios that these two measures correlate strongly with each other and with generalization. Then, we extend the analysis to the deep learning scenario by extensive numerical validations. We study two algorithms, Entropy-SGD and Replicated-SGD, that explicitly include the local entropy in the optimization objective. We devise a training schedule by which we consistently find flatter minima (using both flatness measures), and improve the generalization error for common architectures (e.g. ResNet, EfficientNet).

SEDONA: Search for Decoupled Neural Networks toward Greedy Block-wise Learning

- Myeongjang Pyeon, Jihwan Moon, Taeyoung Hahn, Gunhee Kim
- abstract@open-review(Poster): Backward locking and update locking are well-known sources of inefficiency in backpropagation that prevent from concurrently updating layers. Several works have recently suggested using local error signals to train network blocks asynchronously to overcome these limitations. However, they often require numerous iterations of trial-and-error to find the best configuration for local training, including how to decouple network blocks and which auxiliary networks to use for each block. In this work, we propose a differentiable search algorithm named SEDONA to automate this process. Experimental results show that our algorithm can consistently discover transferable decoupled architectures for VGG and ResNet variants, and significantly outperforms the ones trained with end-to-end backpropagation and other state-of-the-art greedy-leaning methods in CIFAR-10, Tiny-ImageNet and ImageNet.

Coupled Oscillatory Recurrent Neural Network (coRNN): An accurate and (gradient) stable architecture for learning long time dependencies

- T. Konstantin Rusch, Siddhartha Mishra
- abstract@open-review(Oral): Circuits of biological neurons, such as in the functional parts of the brain can be modeled as networks of coupled oscillators. Inspired by the ability of these systems to express a rich set of outputs while keeping (gradients of) state variables bounded, we propose a novel architecture for recurrent neural networks. Our proposed RNN is based on a time-discretization of a system of second-order ordinary differential equations, modeling networks of controlled nonlinear oscillators. We prove precise bounds on the gradients of the hidden states, leading to the mitigation of the exploding and vanishing gradient problem for this RNN. Experiments show that the proposed RNN is comparable in performance to the state of the art on a variety of benchmarks, demonstrating the potential of this architecture to provide stable and accurate RNNs for processing complex sequential data.

not-MIWAE: Deep Generative Modelling with Missing not at Random Data

- Niels Bruun Ipsen, Pierre-Alexandre Mattei, Jes Frellsen
- abstract@open-review(Poster): When a missing process depends on the missing values themselves, it needs to be explicitly modelled and taken into account while doing likelihood-based inference. We present an approach for building and fitting deep latent variable models (DLVMs) in cases where the missing process is dependent on the missing data. Specifically, a deep neural network enables us to flexibly model the conditional distribution of the missingness pattern given the data. This allows for incorporating prior information about the type of missingness (e.g.~self-censoring) into the model. Our inference technique, based on importance-weighted variational inference, involves maximising a lower bound of the joint likelihood. Stochastic gradients of the bound are obtained by using the reparameterisation trick both in latent space and data space. We show on various kinds of data sets and missingness patterns that explicitly modelling the missing process can be invaluable.

Reweighting Augmented Samples by Minimizing the Maximal Expected Loss

- Mingyang Yi, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, Zhi-Ming Ma
- abstract@open-review(Poster): Data augmentation is an effective technique to improve the generalization of deep neural networks. However, previous data augmentation methods usually treat the augmented samples equally without considering their individual impacts on the model. To address this, for the augmented samples from the same training example, we propose to assign different weights to them. We construct the maximal expected loss which is the supremum over any reweighted loss on augmented samples. Inspired by adversarial training, we minimize this maximal expected loss (MMEL) and obtain a simple and interpretable closed-form solution: more attention should be paid to augmented samples with large loss values (i.e., harder examples). Minimizing this maximal expected loss enables the model to perform well under any reweighting strategy. The proposed method can generally be applied on top of any data augmentation methods. Experiments are conducted on both natural language understanding tasks with token-level data augmentation, and image classification tasks with commonly-used image augmentation techniques like random crop and horizontal flip. Empirical results show that the proposed method improves the generalization performance of the model.

Learning to Represent Action Values as a Hypergraph on the Action Vertices

- Arash Tavakoli, Mehdi Fatemi, Petar Kormushev
- abstract@open-review(Poster): Action-value estimation is a critical component of many reinforcement learning (RL) methods whereby sample complexity relies heavily on how fast a good estimator for action value can be learned. By viewing this problem through the lens of representation learning, good representations of both state and action can facilitate action-value estimation. While advances in deep learning have seamlessly driven progress in learning state representations, given the specificity of the notion of agency to RL, little attention has been paid to learning action representations. We conjecture that leveraging the combinatorial structure of multi-dimensional action spaces is a key ingredient for learning good representations of action. To test this, we set forth the action hypergraph networks framework---a class of functions for learning action representations in multi-dimensional discrete action spaces with a structural inductive bias. Using this framework we realise an agent class based on a combination with deep Q-networks, which we dub hypergraph Q-networks. We show the effectiveness of our approach on a myriad of domains: illustrative prediction problems under minimal confounding effects, Atari 2600 games, and discretised physical control benchmarks.

Autoregressive Entity Retrieval

- Nicola De Cao, Gautier Izacard, Sebastian Riedel, Fabio Petroni
- abstract@open-review(Spotlight): Entities are at the center of how we represent and aggregate knowledge. For instance, Encyclopedias such as Wikipedia are structured by entities (e.g., one per Wikipedia article). The ability to retrieve such entities given a query is fundamental for knowledge-intensive tasks such as entity linking and open-domain question answering. One way to understand current approaches is as classifiers among atomic labels, one for each entity. Their weight vectors are dense entity representations produced by encoding entity meta information such as their descriptions. This approach leads to several shortcomings: (i) context and entity affinity is mainly captured through a vector dot product, potentially missing fine-grained interactions between the two; (ii) a large memory footprint is needed to store dense representations when considering large entity sets; (iii) an appropriately hard set of

negative data has to be subsampled at training time. In this work, we propose GENRE, the first system that retrieves entities by generating their unique names, left to right, token-by-token in an autoregressive fashion and conditioned on the context. This enables us to mitigate the aforementioned technical issues since: (i) the autoregressive formulation allows us to directly capture relations between context and entity name, effectively cross encoding both; (ii) the memory footprint is greatly reduced because the parameters of our encoder-decoder architecture scale with vocabulary size, not entity count; (iii) the exact softmax loss can be efficiently computed without the need to subsample negative data. We show the efficacy of the approach, experimenting with more than 20 datasets on entity disambiguation, end-to-end entity linking and document retrieval tasks, achieving new state-of-the-art or very competitive results while using a tiny fraction of the memory footprint of competing systems. Finally, we demonstrate that new entities can be added by simply specifying their unambiguous name. Code and pre-trained models at https://github.com/facebookresearch/GENRE.

Shapley Explanation Networks

- Rui Wang, Xiaoqian Wang, David I. Inouye
- abstract@open-review(Poster): Shapley values have become one of the most popular feature attribution explanation methods. However, most prior work has focused on post-hoc Shapley explanations, which can be computationally demanding due to its exponential time complexity and preclude model regularization based on Shapley explanations during training. Thus, we propose to incorporate Shapley values themselves as latent representations in deep models thereby making Shapley explanations first-class citizens in the modeling paradigm. This intrinsic explanation approach enables layer-wise explanations, explanation regularization of the model during training, and fast explanation computation at test time. We define the Shapley transform that transforms the input into a Shapley representation given a specific function. We operationalize the Shapley transform as a neural network module and construct both shallow and deep networks, called ShapNets, by composing Shapley modules. We prove that our Shallow ShapNets compute the exact Shapley values and our Deep ShapNets maintain the missingness and accuracy properties of Shapley values. We demonstrate on synthetic and real-world datasets that our ShapNets enable layer-wise Shapley explanations, novel Shapley regularizations during training, and fast computation while maintaining reasonable performance. Code is available at https://github.com/inouye-lab/ShapleyExplanationNetworks.

Neural Approximate Sufficient Statistics for Implicit Models

- Yanzhi Chen, Dinghuai Zhang, Michael U. Gutmann, Aaron Courville, Zhanxing Zhu
- abstract@open-review(Spotlight): We consider the fundamental problem of how to automatically construct summary statistics for implicit generative models where the evaluation of the likelihood function is intractable but sampling data from the model is possible. The idea is to frame the task of constructing sufficient statistics as learning mutual information maximizing representations of the data with the help of deep neural networks. The infomax learning procedure does not need to estimate any density or density ratio. We apply our approach to both traditional approximate Bayesian computation and recent neural likelihood methods, boosting their performance on a range of tasks.

NBDT: Neural-Backed Decision Tree

- Alvin Wan, Lisa Dunlap, Daniel Ho, Jihan Yin, Scott Lee, Suzanne Petryk, Sarah Adel Bargal, Joseph E. Gonzalez
- abstract@open-review(Poster): Machine learning applications such as finance and medicine demand accurate and justifiable predictions, barring most deep learning methods from use. In response, previous work combines decision trees with deep learning, yielding models that (1) sacrifice interpretability for accuracy or (2) sacrifice accuracy for interpretability. We forgo this dilemma by jointly improving accuracy and interpretability using Neural-Backed Decision Trees (NBDTs). NBDTs replace a neural network's final linear layer with a differentiable sequence of decisions and a surrogate loss. This forces the model to learn high-level concepts and lessens reliance on highly-uncertain decisions, yielding (1) accuracy: NBDTs match or outperform modern neural networks on CIFAR, ImageNet and better generalize to unseen classes by up to 16%. Furthermore, our surrogate loss improves the original model's accuracy by up to 2%. NBDTs also afford (2) interpretability: improving human trustby clearly identifying model mistakes and assisting in dataset debugging. Code and pretrained NBDTs are at https://github.com/alvinwan/neural-backed-decision-trees.

DiffWave: A Versatile Diffusion Model for Audio Synthesis

- Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, Bryan Catanzaro
- abstract@open-review(Oral): In this work, we propose DiffWave, a versatile diffusion probabilistic model for conditional and unconditional waveform generation. The model is non-autoregressive, and converts the white noise signal into structured waveform through a Markov chain with a constant number of steps at synthesis. It is efficiently trained by optimizing a variant of variational bound on the data likelihood. DiffWave produces high-fidelity audios in different waveform generation tasks, including neural vocoding conditioned on mel spectrogram, class-conditional generation, and unconditional generation. We demonstrate that DiffWave matches a strong WaveNet vocoder in terms of speech quality (MOS: 4.44 versus 4.43), while synthesizing orders of magnitude faster. In particular, it significantly outperforms autoregressive and GAN-based waveform models in the challenging unconditional generation task in terms of audio quality and sample diversity from various automatic and human evaluations.

Denoising Diffusion Implicit Models

- Jiaming Song, Chenlin Meng, Stefano Ermon
- abstract@open-review(Poster): Denoising diffusion probabilistic models (DDPMs) have achieved high quality image generation without adversarial training, yet they require simulating a Markov chain for many steps in order to produce a sample. To accelerate sampling, we present denoising diffusion implicit models (DDIMs), a more efficient class of iterative implicit probabilistic models with the same training procedure as DDPMs. In DDPMs, the generative process is defined as the reverse of a particular Markovian diffusion process. We generalize DDPMs via a class of non-Markovian diffusion processes that lead to the same training objective. These non-Markovian processes can correspond to generative processes that are deterministic, giving rise to implicit models that produce high quality samples much faster. We empirically demonstrate that DDIMs can produce high quality samples $10 \times$ to $50 \times$ faster in terms of wall-clock time compared to DDPMs, allow us to trade off computation for sample quality, perform semantically meaningful image interpolation directly in the latent space, and reconstruct observations with very low error.

Large Scale Image Completion via Co-Modulated Generative Adversarial Networks

- Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I-Chao Chang, Yan Xu
- abstract@open-review(Spotlight): Numerous task-specific variants of conditional generative adversarial networks have been developed for image completion. Yet, a serious limitation remains that all existing algorithms tend to fail when handling large-scale missing regions. To overcome this challenge, we propose a generic new approach that bridges the gap between image-conditional and recent modulated unconditional generative architectures via co-modulation of both conditional and stochastic style representations. Also, due to the lack of good quantitative metrics for image completion, we propose the new Paired/Unpaired Inception Discriminative Score (P-IDS/U-IDS), which robustly measures the perceptual fidelity of inpainted images compared to real images via linear separability in a feature space. Experiments demonstrate superior performance in terms of both quality and diversity over state-of-the-art methods in free-form image completion and easy generalization to image-to-image translation. Code is available at https://github.com/zsyzzsoft/co-mod-gan.

GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding

- Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, Zhifeng Chen
- abstract@open-review(Poster): Neural network scaling has been critical for improving the model quality in many real-world machine learning applications with vast amounts of training data and compute. Although this trend of scaling is affirmed to be a sure-fire approach for better model quality, there are challenges on the path such as the computation cost,ease of programming, and efficient implementation on parallel devices. In this paper we demonstrate conditional computation as a remedy to the above mentioned impediments, and demonstrate its efficacy and utility. We make extensive use of GShard, a module composed of a set of lightweight annotation APIs and an extension to the XLA compiler to enable large scale models with up to trillions of parameters. GShard and conditional computation enable us to scale up multilingual neural machine translation Transformer model with Sparsely-Gated Mixture-of-Experts. We demonstrate that such a giant model with 600 billion parameters can efficiently be trained on 2048 TPU v3 cores in 4 days to achieve far superior quality for translation from 100 languages to English compared to the prior art.

What Should Not Be Contrastive in Contrastive Learning

- Tete Xiao, Xiaolong Wang, Alexei A Efros, Trevor Darrell
- abstract@open-review(Poster): Recent self-supervised contrastive methods have been able to produce impressive transferable visual representations by learning to be invariant to different data augmentations. However, these methods implicitly assume a particular set of representational invariances (e.g., invariance to color), and can perform poorly when a downstream task violates this assumption (e.g., distinguishing red vs. yellow cars). We introduce a contrastive learning framework which does not require prior knowledge of specific, task-dependent invariances. Our model learns to capture varying and invariant factors for visual representations by constructing separate embedding spaces, each of which is invariant to all but one augmentation. We use a multi-head network with a shared backbone which captures information across each augmentation and alone outperforms all baselines on downstream tasks. We further find that the concatenation of the invariant and varying spaces performs best across all tasks we investigate, including coarse-grained, fine-grained, and few-shot downstream classification tasks, and various data corruptions.

Learning Cross-Domain Correspondence for Control with Dynamics Cycle-Consistency

- Qiang Zhang, Tete Xiao, Alexei A Efros, Lerrel Pinto, Xiaolong Wang
- abstract@open-review(Oral): At the heart of many robotics problems is the challenge of learning correspondences across domains. For instance, imitation learning requires obtaining correspondence between humans and robots; sim-to-real requires correspondence between physics simulators and real hardware; transfer learning requires correspondences between different robot environments. In this paper, we propose to learn correspondence across such domains emphasizing on differing modalities (vision and internal state), physics parameters (mass and friction), and morphologies (number of limbs). Importantly, correspondences are learned using unpaired and randomly collected data from the two domains. We propose dynamics cycles that align dynamic robotic behavior across two domains using a cycle consistency constraint. Once this correspondence is found, we can directly transfer the policy trained on one domain to the other, without needing any additional fine-tuning on the second domain. We perform experiments across a variety of problem domains, both in simulation and on real robots. Our framework is able to align uncalibrated monocular video of a real robot arm to dynamic state-action trajectories of a simulated arm without paired data. Video demonstrations of our results are available at: https://sites.google.com/view/cycledynamics .

On Data-Augmentation and Consistency-Based Semi-Supervised Learning

- Atin Ghosh, Alexandre H. Thiery
- abstract@open-review(Poster): Recently proposed consistency-based Semi-Supervised Learning (SSL) methods such as the Pi-model, temporal ensembling, the mean teacher, or the virtual adversarial training, achieve the state of the art results in several SSL tasks. These methods can typically reach performances that are comparable to their fully supervised counterparts while using only a fraction of labelled examples. Despite these methodological advances, the understanding of these methods is still relatively limited. To make progress, we analyse (variations of) the Pi-model in settings where analytically tractable results can be obtained. We establish links with Manifold Tangent Classifiers and demonstrate that the quality of the perturbations is key to obtaining reasonable SSL performances. Furthermore, we propose a simple extension of the Hidden Manifold Model that naturally incorporates data-augmentation schemes and offers a tractable framework for understanding SSL methods.

DDPNOpt: Differential Dynamic Programming Neural Optimizer

- Guan-Horng Liu, Tianrong Chen, Evangelos Theodorou
- abstract@open-review(Spotlight): Interpretation of Deep Neural Networks (DNNs) training as an optimal control problem with nonlinear dynamical systems has received considerable attention recently, yet the algorithmic development remains relatively limited. In this work, we make an attempt along this line by reformulating the training procedure from the trajectory optimization perspective. We first show that most widely-used algorithms for training DNNs can be linked to the Differential Dynamic Programming (DDP), a celebrated second-order method rooted in the Approximate Dynamic Programming. In this vein, we propose a new class of optimizer, DDP Neural Optimizer (DDPNOpt), for training feedforward and convolution networks. DDPNOpt features layer-wise feedback policies which improve convergence and reduce sensitivity to hyper-parameter over existing methods. It outperforms other optimal-control inspired training methods in both convergence and complexity, and is competitive against state-of-the-art first and second order methods. We also observe DDPNOpt has surprising benefit in preventing gradient vanishing. Our work opens up new avenues for principled algorithmic design built upon the optimal control theory.

Diverse Video Generation using a Gaussian Process Trigger

- Gaurav Shrivastava, Abhinav Shrivastava
- abstract@open-review(Poster): Generating future frames given a few context (or past) frames is a challenging task. It requires modeling the temporal coherence of videos as well as multi-modality in terms of diversity in the potential future states. Current variational approaches for video generation tend to marginalize over multi-modal future outcomes. Instead, we propose to explicitly model the multi-modality in the future outcomes and leverage it to sample diverse futures. Our approach, Diverse Video Generator, uses a GP to learn priors on future states given the past and maintains a probability distribution over possible futures given a particular sample. We leverage the changes in this distribution over time to control the sampling of diverse future states by estimating the end of on-going sequences. In particular, we use the variance of GP over the output function space to trigger a change in the action sequence. We achieve state-of-the-art results on diverse future frame generation in terms of reconstruction quality and diversity of the generated sequences.

Monotonic Kronecker-Factored Lattice

- William Taylor Bakst, Nobuyuki Morioka, Erez Louidor
- abstract@open-review(Poster): It is computationally challenging to learn flexible monotonic functions that guarantee model behavior and provide interpretability beyond a few input features, and in a time where minimizing resource use is increasingly important, we must be able to learn such models that are still efficient. In this paper we show how to effectively and efficiently learn such functions using Kronecker-Factored Lattice (KFL), an efficient reparameterization of flexible monotonic lattice regression via Kronecker product. Both computational and storage costs scale linearly in the number of input features, which is a significant improvement over existing methods that grow exponentially. We also show that we can still properly enforce monotonicity and other shape constraints. The KFL function class consists of products of piecewise-linear functions, and the size of the function class can be further increased through ensembling. We prove that the function class of an ensemble of M base KFL models strictly increases as M increases up to a certain threshold. Beyond this threshold, every multilinear interpolated lattice function can be expressed. Our

experimental results demonstrate that KFL trains faster with fewer parameters while still achieving accuracy and evaluation speeds comparable to or better than the baseline methods and preserving monotonicity guarantees on the learned model.

Auto Seg-Loss: Searching Metric Surrogates for Semantic Segmentation

- Hao Li, Chenxin Tao, Xizhou Zhu, Xiaogang Wang, Gao Huang, Jifeng Dai
- abstract@open-review(Poster): Designing proper loss functions is essential in training deep networks. Especially in the field of semantic segmentation, various evaluation metrics have been proposed for diverse scenarios. Despite the success of the widely adopted cross-entropy loss and its variants, the mis-alignment between the loss functions and evaluation metrics degrades the network performance. Meanwhile, manually designing loss functions for each specific metric requires expertise and significant manpower. In this paper, we propose to automate the design of metric-specific loss functions by searching differentiable surrogate losses for each metric. We substitute the non-differentiable operations in the metrics with parameterized functions, and conduct parameter search to optimize the shape of loss surfaces. Two constraints are introduced to regularize the search space and make the search efficient. Extensive experiments on PASCAL VOC and Cityscapes demonstrate that the searched surrogate losses outperform the manually designed loss functions consistently. The searched losses can generalize well to other datasets and networks. Code shall be released at https://github.com/fundamentalvision/Auto-Seg-Loss.

Deformable DETR: Deformable Transformers for End-to-End Object Detection

- Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, Jifeng Dai
- abstract@open-review(Oral): DETR has been recently proposed to eliminate the need for many hand-designed components in object detection while demonstrating good performance. However, it suffers from slow convergence and limited feature spatial resolution, due to the limitation of Transformer attention modules in processing image feature maps. To mitigate these issues, we proposed Deformable DETR, whose attention modules only attend to a small set of key sampling points around a reference. Deformable DETR can achieve better performance than DETR (especially on small objects) with 10$\times$ less training epochs. Extensive experiments on the COCO benchmark demonstrate the effectiveness of our approach. Code is released at https://github.com/fundamentalvision/Deformable-DETR.

Combining Physics and Machine Learning for Network Flow Estimation

- Arlei Lopes da Silva, Furkan Kocayusufoglu, Saber Jafarpour, Francesco Bullo, Ananthram Swami, Ambuj Singh
- abstract@open-review(Poster): The flow estimation problem consists of predicting missing edge flows in a network (e.g., traffic, power, and water) based on partial observations. These missing flows depend both on the underlying \textit{physics} (edge features and a flow conservation law) as well as the observed edge flows. This paper introduces an optimization framework for computing missing edge flows and solves the problem using bilevel optimization and deep learning. More specifically, we learn regularizers that depend on edge features (e.g., number of lanes in a road, the resistance of a power line) using neural networks. Empirical results show that our method accurately predicts missing flows, outperforming the best baseline, and is able to capture relevant physical properties in traffic and power networks.

NeMo: Neural Mesh Models of Contrastive Features for Robust 3D Pose Estimation

- Angtian Wang, Adam Kortylewski, Alan Yuille
- abstract@open-review(Poster): 3D pose estimation is a challenging but important task in computer vision. In this work, we show that standard deep learning approaches to 3D pose estimation are not robust to partial occlusion. Inspired by the robustness of generative vision models to partial occlusion, we propose to integrate deep neural networks with 3D generative representations of objects into a unified neural architecture that we term NeMo. In particular, NeMo learns a generative model of neural feature activations at each vertex on a dense 3D mesh. Using differentiable rendering we estimate the 3D object pose by minimizing the reconstruction error between NeMo and the feature representation of the target image. To avoid local optima in the reconstruction loss, we train the feature extractor to maximize the distance between the individual feature representations on the mesh using contrastive learning. Our extensive experiments on PASCAL3D+, occluded-PASCAL3D+ and ObjectNet3D show that NeMo is much more robust to partial occlusion compared to standard deep networks, while retaining competitive performance on non-occluded data. Interestingly, our experiments also show that NeMo performs reasonably well even when the mesh representation only crudely approximates the true object geometry with a cuboid, hence revealing that the detailed 3D geometry is not needed for accurate 3D pose estimation.

How to Find Your Friendly Neighborhood: Graph Attention Design with Self-Supervision

- Dongkwan Kim, Alice Oh
- abstract@open-review(Poster): Attention mechanism in graph neural networks is designed to assign larger weights to important neighbor nodes for better representation. However, what graph attention learns is not understood well, particularly when graphs are noisy. In this paper, we propose a self-supervised graph attention network (SuperGAT), an improved graph attention model for noisy graphs. Specifically, we exploit two attention forms compatible with a self-supervised task to predict edges, whose presence and absence contain the inherent information about the importance of the relationships between nodes. By encoding edges, SuperGAT learns more expressive attention in distinguishing mislinked neighbors. We find two graph characteristics influence the effectiveness of attention forms and self-supervision: homophily and average degree. Thus, our recipe provides guidance on which attention design to use when those two graph characteristics are known. Our experiment on 17 real-world datasets demonstrates that our recipe generalizes across 15 datasets of them, and our models designed by recipe show improved performance over baselines.

MiCE: Mixture of Contrastive Experts for Unsupervised Image Clustering

- Tsung Wei Tsai, Chongxuan Li, Jun Zhu
- abstract@open-review(Poster): We present Mixture of Contrastive Experts (MiCE), a unified probabilistic clustering framework that simultaneously exploits the discriminative representations learned by contrastive learning and the semantic structures captured by a latent mixture model. Motivated by the mixture of experts, MiCE employs a gating function to partition an unlabeled dataset into subsets according to the latent semantics and multiple experts to discriminate distinct subsets of instances assigned to them in a contrastive learning manner. To solve the nontrivial inference and learning problems caused by the latent variables, we further develop a scalable variant of the Expectation-Maximization (EM) algorithm for MiCE and provide proof of the convergence. Empirically, we evaluate the clustering performance of MiCE on four widely adopted natural image datasets. MiCE achieves significantly better results than various previous methods and a strong contrastive learning baseline.

Anytime Sampling for Autoregressive Models via Ordered Autoencoding

- Yilun Xu, Yang Song, Sahaj Garg, Linyuan Gong, Rui Shu, Aditya Grover, Stefano Ermon
- abstract@open-review(Poster): Autoregressive models are widely used for tasks such as image and audio generation. The sampling process of these models, however, does not allow interruptions and cannot adapt to real-time computational resources. This challenge impedes the deployment of powerful autoregressive models, which involve a slow sampling process that is sequential in nature and typically scales linearly with respect to the data dimension. To address this difficulty, we propose a new family of autoregressive models that enables anytime sampling. Inspired by Principal Component Analysis, we learn a structured representation space where dimensions are ordered based on their importance with respect to reconstruction. Using an autoregressive

model in this latent space, we trade off sample quality for computational efficiency by truncating the generation process before decoding into the original data space. Experimentally, we demonstrate in several image and audio generation tasks that sample quality degrades gracefully as we reduce the computational budget for sampling. The approach suffers almost no loss in sample quality (measured by FID) using only 60\% to 80\% of all latent dimensions for image data. Code is available at https://github.com/Newbeeer/Anytime-Auto-Regressive-Model.

Initialization and Regularization of Factorized Neural Layers

- Mikhail Khodak, Neil A. Tenenholtz, Lester Mackey, Nicolo Fusi
- abstract@open-review(Poster): Factorized layers—operations parameterized by products of two or more matrices—occur in a variety of deep learning contexts, including compressed model training, certain types of knowledge distillation, and multi-head self-attention architectures. We study how to initialize and regularize deep nets containing such layers, examining two simple, understudied schemes, spectral initialization and Frobenius decay, for improving their performance. The guiding insight is to design optimization routines for these networks that are as close as possible to that of their well-tuned, non-decomposed counterparts; we back this intuition with an analysis of how the initialization and regularization schemes impact training with gradient descent, drawing on modern attempts to understand the interplay of weight-decay and batch-normalization. Empirically, we highlight the benefits of spectral initialization and Frobenius decay across a variety of settings. In model compression, we show that they enable low-rank methods to significantly outperform both unstructured sparsity and tensor methods on the task of training low-memory residual networks; analogs of the schemes also improve the performance of tensor decomposition techniques. For knowledge distillation, Frobenius decay enables a simple, overcomplete baseline that yields a compact model from over-parameterized training without requiring retraining with or pruning a teacher network. Finally, we show how both schemes applied to multi-head attention lead to improved performance on both translation and unsupervised pre-training.

CO2: Consistent Contrast for Unsupervised Visual Representation Learning

- Chen Wei, Huiyu Wang, Wei Shen, Alan Yuille
- abstract@open-review(Poster): Contrastive learning has recently been a core for unsupervised visual representation learning. Without human annotation, the common practice is to perform an instance discrimination task: Given a query image crop, label crops from the same image as positives, and crops from other randomly sampled images as negatives. An important limitation of this label assignment is that it can not reflect the heterogeneous similarity of the query crop to crops from other images, but regarding them as equally negative. To address this issue, inspired by consistency regularization in semi-supervised learning, we propose Consistent Contrast (CO2), which introduces a consistency term into unsupervised contrastive learning framework. The consistency term takes the similarity of the query crop to crops from other images as unlabeled, and the corresponding similarity of a positive crop as a pseudo label. It then encourages consistency between these two similarities. Empirically, CO2 improves Momentum Contrast (MoCo) by 2.9% top-1 accuracy on ImageNet linear protocol, 3.8% and 1.1% top-5 accuracy on 1% and 10% labeled semi-supervised settings. It also transfers to image classification, object detection, and semantic segmentation on PASCAL VOC. This shows that CO2 learns better visual representations for downstream tasks.

Geometry-Aware Gradient Algorithms for Neural Architecture Search

- Liam Li, Mikhail Khodak, Nina Balcan, Ameet Talwalkar
- abstract@open-review(Spotlight): Recent state-of-the-art methods for neural architecture search (NAS) exploit gradient-based optimization by relaxing the problem into continuous optimization over architectures and shared-weights, a noisy process that remains poorly understood. We argue for the study of single-level empirical risk minimization to understand NAS with weight-sharing, reducing the design of NAS methods to devising optimizers and regularizers that can quickly obtain high-quality solutions to this problem. Invoking the theory of mirror descent, we present a geometry-aware framework that exploits the underlying structure of this optimization to return sparse architectural parameters, leading to simple yet novel algorithms that enjoy fast convergence guarantees and achieve state-of-the-art accuracy on the latest NAS benchmarks in computer vision. Notably, we exceed the best published results for both CIFAR and ImageNet on both the DARTS search space and NAS-Bench-201; on the latter we achieve near-oracle-optimal performance on CIFAR-10 and CIFAR-100. Together, our theory and experiments demonstrate a principled way to co-design optimizers and continuous relaxations of discrete NAS search spaces.

Beyond Fully-Connected Layers with Quaternions: Parameterization of Hypercomplex Multiplications with $1/n$ Parameters

- Aston Zhang, Yi Tay, SHUAI Zhang, Alvin Chan, Anh Tuan Luu, Siu Hui, Jie Fu
- abstract@open-review(Spotlight): Recent works have demonstrated reasonable success of representation learning in hypercomplex space. Specifically, “fully-connected layers with quaternions” (quaternions are 4D hypercomplex numbers), which replace real-valued matrix multiplications in fully-connected layers with Hamilton products of quaternions, both enjoy parameter savings with only 1/4 learnable parameters and achieve comparable performance in various applications. However, one key caveat is that hypercomplex space only exists at very few predefined dimensions (4D, 8D, and 16D). This restricts the flexibility of models that leverage hypercomplex multiplications. To this end, we propose parameterizing hypercomplex multiplications, allowing models to learn multiplication rules from data regardless of whether such rules are predefined. As a result, our method not only subsumes the Hamilton product, but also learns to operate on any arbitrary nD hypercomplex space, providing more architectural flexibility using arbitrarily $1/n$ learnable parameters compared with the fully-connected layer counterpart. Experiments of applications to the LSTM and transformer models on natural language inference, machine translation, text style transfer, and subject verb agreement demonstrate architectural flexibility and effectiveness of the proposed approach.

Tent: Fully Test-Time Adaptation by Entropy Minimization

- Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, Trevor Darrell
- abstract@open-review(Spotlight): A model must adapt itself to generalize to new and different data during testing. In this setting of fully test-time adaptation the model has only the test data and its own parameters. We propose to adapt by test entropy minimization (tent): we optimize the model for confidence as measured by the entropy of its predictions. Our method estimates normalization statistics and optimizes channel-wise affine transformations to update online on each batch. Tent reduces generalization error for image classification on corrupted ImageNet and CIFAR-10/100 and reaches a new state-of-the-art error on ImageNet-C. Tent handles source-free domain adaptation on digit recognition from SVHN to MNIST/MNIST-M/USPS, on semantic segmentation from GTA to Cityscapes, and on the VisDA-C benchmark. These results are achieved in one epoch of test-time optimization without altering training.

DC3: A learning method for optimization with hard constraints

- Priya L. Donti, David Rolnick, J Zico Kolter
- abstract@open-review(Poster): Large optimization problems with hard constraints arise in many settings, yet classical solvers are often prohibitively slow, motivating the use of deep networks as cheap "approximate solvers." Unfortunately, naive deep learning approaches typically cannot enforce the hard constraints of such problems, leading to infeasible solutions. In this work, we present Deep Constraint Completion and Correction (DC3), an algorithm to address this challenge. Specifically, this method enforces feasibility via a differentiable procedure, which implicitly completes partial solutions to satisfy equality constraints and unrolls gradient-based corrections to satisfy inequality constraints. We demonstrate the effectiveness of DC3 in both synthetic optimization tasks and the real-world setting of AC optimal power flow, where hard constraints encode the physics of the electrical grid. In both cases, DC3 achieves near-optimal objective values while preserving feasibility.

Enforcing robust control guarantees within neural network policies

- Priya L. Donti, Melrose Roderick, Mahyar Fazlyab, J Zico Kolter
- abstract@open-review(Poster): When designing controllers for safety-critical systems, practitioners often face a challenging tradeoff between robustness and performance. While robust control methods provide rigorous guarantees on system stability under certain worst-case disturbances, they often yield simple controllers that perform poorly in the average (non-worst) case. In contrast, nonlinear control methods trained using deep learning have achieved state-of-the-art performance on many control tasks, but often lack robustness guarantees. In this paper, we propose a technique that combines the strengths of these two approaches: constructing a generic nonlinear control policy class, parameterized by neural networks, that nonetheless enforces the same provable robustness criteria as robust control. Specifically, our approach entails integrating custom convex-optimization-based projection layers into a neural network-based policy. We demonstrate the power of this approach on several domains, improving in average-case performance over existing robust control methods and in worst-case stability over (non-robust) deep RL methods.

Neural Topic Model via Optimal Transport

- He Zhao, Dinh Phung, Viet Huynh, Trung Le, Wray Buntine
- abstract@open-review(Spotlight): Recently, Neural Topic Models (NTMs) inspired by variational autoencoders have obtained increasingly research interest due to their promising results on text analysis. However, it is usually hard for existing NTMs to achieve good document representation and coherent/diverse topics at the same time. Moreover, they often degrade their performance severely on short documents. The requirement of reparameterisation could also comprise their training quality and model flexibility. To address these shortcomings, we present a new neural topic model via the theory of optimal transport (OT). Specifically, we propose to learn the topic distribution of a document by directly minimising its OT distance to the document's word distributions. Importantly, the cost matrix of the OT distance models the weights between topics and words, which is constructed by the distances between topics and words in an embedding space. Our proposed model can be trained efficiently with a differentiable loss. Extensive experiments show that our framework significantly outperforms the state-of-the-art NTMs on discovering more coherent and diverse topics and deriving better document representations for both regular and short texts.

Robust and Generalizable Visual Representation Learning via Random Convolutions

- Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, Marc Niethammer
- abstract@open-review(Poster): While successful for various computer vision tasks, deep neural networks have shown to be vulnerable to texture style shifts and small perturbations to which humans are robust. In this work, we show that the robustness of neural networks can be greatly improved through the use of random convolutions as data augmentation. Random convolutions are approximately shape-preserving and may distort local textures. Intuitively, randomized convolutions create an infinite number of new domains with similar global shapes but random local texture. Therefore, we explore using outputs of multi-scale random convolutions as new images or mixing them with the original images during training. When applying a network trained with our approach to unseen domains, our method consistently improves the performance on domain generalization benchmarks and is scalable to ImageNet. In particular, in the challenging scenario of generalizing to the sketch domain in PACS and to ImageNet-Sketch, our method outperforms state-of-art methods by a large margin. More interestingly, our method can benefit downstream tasks by providing a more robust pretrained visual representation.

WrapNet: Neural Net Inference with Ultra-Low-Precision Arithmetic

- Renkun Ni, Hong-min Chu, Oscar Castaneda, Ping-yeh Chiang, Christoph Studer, Tom Goldstein
- abstract@open-review(Poster): Low-precision neural networks represent both weights and activations with few bits, drastically reducing the cost of multiplications. Meanwhile, these products are accumulated using high-precision (typically 32-bit) additions. Additions dominate the arithmetic complexity of inference in quantized (e.g., binary) nets, and high precision is needed to avoid overflow. To further optimize inference, we propose WrapNet, an architecture that adapts neural networks to use low-precision (8-bit) additions while achieving classification accuracy comparable to their 32-bit counterparts. We achieve resilience to low-precision accumulation by inserting a cyclic activation layer that makes results invariant to overflow. We demonstrate the efficacy of our approach using both software and hardware platforms.

Rank the Episodes: A Simple Approach for Exploration in Procedurally-Generated Environments

- Daochen Zha, Wenye Ma, Lei Yuan, Xia Hu, Ji Liu
- abstract@open-review(Poster): Exploration under sparse reward is a long-standing challenge of model-free reinforcement learning. The state-of-the-art methods address this challenge by introducing intrinsic rewards to encourage exploration in novel states or uncertain environment dynamics. Unfortunately, methods based on intrinsic rewards often fall short in procedurally-generated environments, where a different environment is generated in each episode so that the agent is not likely to visit the same state more than once. Motivated by how humans distinguish good exploration behaviors by looking into the entire episode, we introduce RAPID, a simple yet effective episode-level exploration method for procedurally-generated environments. RAPID regards each episode as a whole and gives an episodic exploration score from both per-episode and long-term views. Those highly scored episodes are treated as good exploration behaviors and are stored in a small ranking buffer. The agent then imitates the episodes in the buffer to reproduce the past good exploration behaviors. We demonstrate our method on several procedurally-generated MiniGrid environments, a first-person-view 3D Maze navigation task from MiniWorld, and several sparse MuJoCo tasks. The results show that RAPID significantly outperforms the state-of-the-art intrinsic reward strategies in terms of sample efficiency and final performance. The code is available at https://github.com/daochenzha/rapid

Stochastic Security: Adversarial Defense Using Long-Run Dynamics of Energy-Based Models

- Mitch Hill, Jonathan Craig Mitchell, Song-Chun Zhu
- abstract@open-review(Poster): The vulnerability of deep networks to adversarial attacks is a central problem for deep learning from the perspective of both cognition and security. The current most successful defense method is to train a classifier using adversarial images created during learning. Another defense approach involves transformation or purification of the original input to remove adversarial signals before the image is classified. We focus on defending naturally-trained classifiers using Markov Chain Monte Carlo (MCMC) sampling with an Energy-Based Model (EBM) for adversarial purification. In contrast to adversarial training, our approach is intended to secure highly vulnerable pre-existing classifiers. To our knowledge, no prior defensive transformation is capable of securing naturally-trained classifiers, and our method is the first to validate a post-training defense approach that is distinct from current successful defenses which modify classifier training.

The memoryless behavior of long-run MCMC sampling will eventually remove adversarial signals, while metastable behavior preserves consistent appearance of MCMC samples after many steps to allow accurate long-run prediction. Balancing these factors can lead to effective purification and robust classification. We evaluate adversarial defense with an EBM using the strongest known attacks against purification. Our contributions are 1) an improved method for training EBM's with realistic long-run MCMC samples for effective purification, 2) an Expectation-Over-Transformation (EOT) defense that resolves ambiguities for evaluating stochastic defenses and from which the EOT attack naturally follows, and 3) state-of-the-art adversarial defense for naturally-trained classifiers and competitive defense compared to adversarial training on CIFAR-10, SVHN, and CIFAR-100. Our code and pre-trained models are available at https://github.com/point0bar1/ebm-defense.

Nearest Neighbor Machine Translation

- Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, Mike Lewis
- abstract@open-review(Poster): We introduce k-nearest-neighbor machine translation (kNN-MT), which predicts tokens with a nearest-neighbor classifier over a large datastore of cached examples, using representations from a neural translation model for similarity search. This approach requires no additional training and scales to give the decoder direct access to billions of examples at test time, resulting in a highly expressive model that consistently improves performance across many settings. Simply adding nearest-neighbor search improves a state-of-the-art German-English translation model by 1.5 BLEU. kNN-MT allows a single model to be adapted to diverse domains by using a domain-specific datastore, improving results by an average of 9.2 BLEU over zero-shot transfer, and achieving new state-of-the-art results---without training on these domains. A massively multilingual model can also be specialized for particular language pairs, with improvements of 3 BLEU for translating from English into German and Chinese. Qualitatively, kNN-MT is easily interpretable; it combines source and target context to retrieve highly relevant examples.

Learning Hyperbolic Representations of Topological Features

- Panagiotis Kyriakis, Iordanis Fostiropoulos, Paul Bogdan
- abstract@open-review(Poster): Learning task-specific representations of persistence diagrams is an important problem in topological data analysis and machine learning. However, current state of the art methods are restricted in terms of their expressivity as they are focused on Euclidean representations. Persistence diagrams often contain features of infinite persistence (i.e., essential features) and Euclidean spaces shrink their importance relative to non-essential features because they cannot assign infinite distance to finite points. To deal with this issue, we propose a method to learn representations of persistence diagrams on hyperbolic spaces, more specifically on the Poincare ball. By representing features of infinite persistence infinitesimally close to the boundary of the ball, their distance to non-essential features approaches infinity, thereby their relative importance is preserved. This is achieved without utilizing extremely high values for the learnable parameters, thus the representation can be fed into downstream optimization methods and trained efficiently in an end-to-end fashion. We present experimental results on graph and image classification tasks and show that the performance of our method is on par with or exceeds the performance of other state of the art methods.

A Panda? No, It's a Sloth: Slowdown Attacks on Adaptive Multi-Exit Neural Network Inference

- Sanghyun Hong, Yigitcan Kaya, Ionuț-Vlad Modoranu, Tudor Dumitras
- abstract@open-review(Spotlight): Recent increases in the computational demands of deep neural networks (DNNs), combined with the observation that most input samples require only simple models, have sparked interest in input-adaptive multi-exit architectures, such as MSDNets or Shallow-Deep Networks. These architectures enable faster inferences and could bring DNNs to low-power devices, e.g., in the Internet of Things (IoT). However, it is unknown if the computational savings provided by this approach are robust against adversarial pressure. In particular, an adversary may aim to slowdown adaptive DNNs by increasing their average inference time—a threat analogous to the denial-of-service attacks from the Internet. In this paper, we conduct a systematic evaluation of this threat by experimenting with three generic multi-exit DNNs (based on VGG16, MobileNet, and ResNet56) and a custom multi-exit architecture, on two popular image classification benchmarks (CIFAR-10 and Tiny ImageNet). To this end, we show that adversarial example-crafting techniques can be modified to cause slowdown, and we propose a metric for comparing their impact on different architectures. We show that a slowdown attack reduces the efficacy of multi-exit DNNs by 90–100%, and it amplifies the latency by 1.5–5× in a typical IoT deployment. We also show that it is possible to craft universal, reusable perturbations and that the attack can be effective in realistic black-box scenarios, where the attacker has limited knowledge about the victim. Finally, we show that adversarial training provides limited protection against slowdowns. These results suggest that further research is needed for defending multi-exit architectures against this emerging threat. Our code is available at https://github.com/sanghyun-hong/deepsloth.

Selectivity considered harmful: evaluating the causal impact of class selectivity in DNNs

- Matthew L Leavitt, Ari S. Morcos
- abstract@open-review(Poster): The properties of individual neurons are often analyzed in order to understand the biological and artificial neural networks in which they're embedded. Class selectivity—typically defined as how different a neuron's responses are across different classes of stimuli or data samples—is commonly used for this purpose. However, it remains an open question whether it is necessary and/or sufficient for deep neural networks (DNNs) to learn class selectivity in individual units. We investigated the causal impact of class selectivity on network function by directly regularizing for or against class selectivity. Using this regularizer to reduce class selectivity across units in convolutional neural networks increased test accuracy by over 2% in ResNet18 and 1% in ResNet50 trained on Tiny ImageNet. For ResNet20 trained on CIFAR10 we could reduce class selectivity by a factor of 2.5 with no impact on test accuracy, and reduce it nearly to zero with only a small (~2%) drop in test accuracy. In contrast, regularizing to increase class selectivity significantly decreased test accuracy across all models and datasets. These results indicate that class selectivity in individual units is neither sufficient nor strictly necessary, and can even impair DNN performance. They also encourage caution when focusing on the properties of single units as representative of the mechanisms by which DNNs function.

Integrating Categorical Semantics into Unsupervised Domain Translation

- Samuel Lavoie-Marchildon, Faruk Ahmed, Aaron Courville
- abstract@open-review(Poster): While unsupervised domain translation (UDT) has seen a lot of success recently, we argue that mediating its translation via categorical semantic features could broaden its applicability. In particular, we demonstrate that categorical semantics improves the translation between perceptually different domains sharing multiple object categories. We propose a method to learn, in an unsupervised manner, categorical semantic features (such as object labels) that are invariant of the source and target domains. We show that conditioning the style encoder of unsupervised domain translation methods on the learned categorical semantics leads to a translation preserving the digits on MNIST$\leftrightarrow$SVHN and to a more realistic stylization on Sketches$\to$Reals.

Are Neural Rankers still Outperformed by Gradient Boosted Decision Trees?

- Zhen Qin, Le Yan, Honglei Zhuang, Yi Tay, Rama Kumar Pasumarthi, Xuanhui Wang, Michael Bendersky, Marc Najork
- abstract@open-review(Spotlight): Despite the success of neural models on many major machine learning problems, their effectiveness on traditional Learning-to-Rank (LTR) problems is still not widely acknowledged. We first validate this concern by showing that most recent neural LTR models are, by a large margin, inferior to the best publicly available Gradient Boosted Decision Trees (GBDT) in terms of their reported ranking accuracy on benchmark datasets. This unfortunately was somehow overlooked in recent neural LTR papers. We then investigate why existing neural LTR models under-perform and identify several of their weaknesses. Furthermore, we propose a unified framework comprising of counter strategies to ameliorate the existing weaknesses of neural models. Our models are the first to be able to perform equally well, comparing with the best tree-based baseline, while outperforming recently published neural LTR models by a large margin. Our results can also serve as a benchmark to facilitate future improvement of neural LTR models.

Zero-shot Synthesis with Group-Supervised Learning

- Yunhao Ge, Sami Abu-El-Haija, Gan Xin, Laurent Itti
- abstract@open-review(Poster): Visual cognition of primates is superior to that of artificial neural networks in its ability to “envision” a visual object, even a newly-introduced one, in different attributes including pose, position, color, texture, etc. To aid neural networks to envision objects with different attributes, we propose a family of objective functions, expressed on groups of examples, as a novel learning framework that we term Group-Supervised

Learning (GSL). GSL allows us to decompose inputs into a disentangled representation with swappable components, that can be recombined to synthesize new samples. For instance, images of red boats & blue cars can be decomposed and recombined to synthesize novel images of red cars. We propose an implementation based on auto-encoder, termed group-supervised zero-shot synthesis network (GZS-Net) trained with our learning framework, that can produce a high-quality red car even if no such example is witnessed during training. We test our model and learning framework on existing benchmarks, in addition to a new dataset that we open-source. We qualitatively and quantitatively demonstrate that GZS-Net trained with GSL outperforms state-of-the-art methods

Learning Generalizable Visual Representations via Interactive Gameplay

- Luca Weihs, Aniruddha Kembhavi, Kiana Ehsani, Sarah M Pratt, Winson Han, Alvaro Herrasti, Eric Kolve, Dustin Schwenk, Roozbeh Mottaghi, Ali Farhadi
- abstract@open-review(Oral): A growing body of research suggests that embodied gameplay, prevalent not just in human cultures but across a variety of animal species including turtles and ravens, is critical in developing the neural flexibility for creative problem solving, decision making, and socialization. Comparatively little is known regarding the impact of embodied gameplay upon artificial agents. While recent work has produced agents proficient in abstract games, these environments are far removed the real world and thus these agents can provide little insight into the advantages of embodied play. Hiding games, such as hide-and-seek, played universally, provide a rich ground for studying the impact of embodied gameplay on representation learning in the context of perspective taking, secret keeping, and false belief understanding. Here we are the first to show that embodied adversarial reinforcement learning agents playing Cache, a variant of hide-and-seek, in a high fidelity, interactive, environment, learn generalizable representations of their observations encoding information such as object permanence, free space, and containment. Moving closer to biologically motivated learning strategies, our agents' representations, enhanced by intentionality and memory, are developed through interaction and play. These results serve as a model for studying how facets of vision develop through interaction, provide an experimental framework for assessing what is learned by artificial agents, and demonstrates the value of moving from large, static, datasets towards experiential, interactive, representation learning.

VA-RED^2: Video Adaptive Redundancy Reduction

- Bowen Pan, Rameswar Panda, Camilo Luciano Fosco, Chung-Ching Lin, Alex J Andonian, Yue Meng, Kate Saenko, Aude Oliva, Rogerio Feris
- abstract@open-review(Poster): Performing inference on deep learning models for videos remains a challenge due to the large amount of computational resources required to achieve robust recognition. An inherent property of real-world videos is the high correlation of information across frames which can translate into redundancy in either temporal or spatial feature maps of the models, or both. The type of redundant features depends on the dynamics and type of events in the video: static videos have more temporal redundancy while videos focusing on objects tend to have more channel redundancy. Here we present a redundancy reduction framework, termed VA-RED^2, which is input-dependent. Specifically, our VA-RED^2 framework uses an input-dependent policy to decide how many features need to be computed for temporal and channel dimensions. To keep the capacity of the original model, after fully computing the necessary features, we reconstruct the remaining redundant features from those using cheap linear operations. We learn the adaptive policy jointly with the network weights in a differentiable way with a shared-weight mechanism, making it highly efficient. Extensive experiments on multiple video datasets and different visual tasks show that our framework achieves $20\% - 40\%$ reduction in computation (FLOPs) when compared to state-of-the-art methods without any performance loss. Project page: http://people.csail.mit.edu/bpan/va-red/.

BERTology Meets Biology: Interpreting Attention in Protein Language Models

- Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, richard socher, Nazneen Rajani
- abstract@open-review(Poster): Transformer architectures have proven to learn useful representations for protein classification and generation tasks. However, these representations present challenges in interpretability. In this work, we demonstrate a set of methods for analyzing protein Transformer models through the lens of attention. We show that attention: (1) captures the folding structure of proteins, connecting amino acids that are far apart in the underlying sequence, but spatially close in the three-dimensional structure, (2) targets binding sites, a key functional component of proteins, and (3) focuses on progressively more complex biophysical properties with increasing layer depth. We find this behavior to be consistent across three Transformer architectures (BERT, ALBERT, XLNet) and two distinct protein datasets. We also present a three-dimensional visualization of the interaction between attention and protein structure. Code for visualization and analysis is available at https://github.com/salesforce/provis.

Trajectory Prediction using Equivariant Continuous Convolution

- Robin Walters, Jinxi Li, Rose Yu
- abstract@open-review(Poster): Trajectory prediction is a critical part of many AI applications, for example, the safe operation of autonomous vehicles. However, current methods are prone to making inconsistent and physically unrealistic predictions. We leverage insights from fluid dynamics to overcome this limitation by considering internal symmetry in real-world trajectories. We propose a novel model, Equivariant Continous COnvolution (ECCO) for improved trajectory prediction. ECCO uses rotationally-equivariant continuous convolutions to embed the symmetries of the system. On both vehicle and pedestrian trajectory datasets, ECCO attains competitive accuracy with significantly fewer parameters. It is also more sample efficient, generalizing automatically from few data points in any orientation. Lastly, ECCO improves generalization with equivariance, resulting in more physically consistent predictions. Our method provides a fresh perspective towards increasing trust and transparency in deep learning models. Our code and data can be found at https://github.com/Rose-STL-Lab/ECCO.

Unsupervised Meta-Learning through Latent-Space Interpolation in Generative Models

- Siavash Khodadadeh, Sharare Zehtabian, Saeed Vahidian, Weijia Wang, Bill Lin, Ladislau Boloni
- abstract@open-review(Poster): Several recently proposed unsupervised meta-learning approaches rely on synthetic meta-tasks created using techniques such as random selection, clustering and/or augmentation. In this work, we describe a novel approach that generates meta-tasks using generative models. The proposed family of algorithms generate pairs of in-class and out-of-class samples from the latent space in a principled way, allowing us to create synthetic classes forming the training and validation data of a meta-task. We find that the proposed approach, LAtent Space Interpolation Unsupervised Meta-learning (LASIUM), outperforms or is competitive with current unsupervised learning baselines on few-shot classification tasks on the most widely used benchmark datasets.

Heteroskedastic and Imbalanced Deep Learning with Adaptive Regularization

- Kaidi Cao, Yining Chen, Junwei Lu, Nikos Arechiga, Adrien Gaidon, Tengyu Ma
- abstract@open-review(Poster): Real-world large-scale datasets are heteroskedastic and imbalanced --- labels have varying levels of uncertainty and label distributions are long-tailed. Heteroskedasticity and imbalance challenge deep learning algorithms due to the difficulty of distinguishing among mislabeled, ambiguous, and rare examples. Addressing heteroskedasticity and imbalance simultaneously is under-explored. We propose a data-dependent regularization technique for heteroskedastic datasets that regularizes different regions of the input space differently. Inspired by the theoretical derivation of the optimal regularization strength in a one-dimensional nonparametric classification setting, our approach adaptively regularizes the data points in higher-uncertainty, lower-density regions more heavily. We test our method on several benchmark tasks, including a real-world heteroskedastic and imbalanced dataset, WebVision. Our experiments corroborate our theory and demonstrate a significant improvement over other methods in noise-robust deep learning.

Implicit Under-Parameterization Inhibits Data-Efficient Deep Reinforcement Learning

- Aviral Kumar, Rishabh Agarwal, Dibya Ghosh, Sergey Levine
- abstract@open-review(Poster): We identify an implicit under-parameterization phenomenon in value-based deep RL methods that use bootstrapping: when value functions, approximated using deep neural networks, are trained with gradient descent using iterated regression onto target values generated by previous instances of the value network, more gradient updates decrease the expressivity of the current value network. We char- acterize this loss of expressivity via a drop in the rank of the learned value net- work features, and show that this typically corresponds to a performance drop. We demonstrate this phenomenon on Atari and Gym benchmarks, in both offline and online RL settings. We formally analyze this phenomenon and show that it results from a pathological interaction between bootstrapping and gradient-based optimization. We further show that mitigating implicit under-parameterization by controlling rank collapse can improve performance.

Bowtie Networks: Generative Modeling for Joint Few-Shot Recognition and Novel-View Synthesis

- Zhipeng Bao, Yu-Xiong Wang, Martial Hebert
- abstract@open-review(Poster): We propose a novel task of joint few-shot recognition and novel-view synthesis: given only one or few images of a novel object from arbitrary views with only category annotation, we aim to simultaneously learn an object classifier and generate images of that type of object from new viewpoints. While existing work copes with two or more tasks mainly by multi-task learning of shareable feature representations, we take a different perspective. We focus on the interaction and cooperation between a generative model and a discriminative model, in a way that facilitates knowledge to flow across tasks in complementary directions. To this end, we propose bowtie networks that jointly learn 3D geometric and semantic representations with a feedback loop. Experimental evaluation on challenging fine-grained recognition datasets demonstrates that our synthesized images are realistic from multiple viewpoints and significantly improve recognition performance as ways of data augmentation, especially in the low-data regime.

Saliency is a Possible Red Herring When Diagnosing Poor Generalization

- Joseph D Viviano, Becks Simpson, Francis Dutil, Yoshua Bengio, Joseph Paul Cohen
- abstract@open-review(Poster): Poor generalization is one symptom of models that learn to predict target variables using spuriously-correlated image features present only in the training distribution instead of the true image features that denote a class. It is often thought that this can be diagnosed visually using attribution (aka saliency) maps. We study if this assumption is correct. In some prediction tasks, such as for medical images, one may have some images with masks drawn by a human expert, indicating a region of the image containing relevant information to make the prediction. We study multiple methods that take advantage of such auxiliary labels, by training networks to ignore distracting features which may be found outside of the region of interest. This mask information is only used during training and has an impact on generalization accuracy depending on the severity of the shift between the training and test distributions. Surprisingly, while these methods improve generalization performance in the presence of a covariate shift, there is no strong correspondence between the correction of attribution towards the features a human expert have labelled as important and generalization performance. These results suggest that the root cause of poor generalization may not always be spatially defined, and raise questions about the utility of masks as 'attribution priors' as well as saliency maps for explainable predictions.

What Can You Learn From Your Muscles? Learning Visual Representation from Human Interactions

- Kiana Ehsani, Daniel Gordon, Thomas Hai Dang Nguyen, Roozbeh Mottaghi, Ali Farhadi
- abstract@open-review(Poster): Learning effective representations of visual data that generalize to a variety of downstream tasks has been a long quest for computer vision. Most representation learning approaches rely solely on visual data such as images or videos. In this paper, we explore a novel approach, where we use human interaction and attention cues to investigate whether we can learn better representations compared to visual-only representations. For this study, we collect a dataset of human interactions capturing body part movements and gaze in their daily lives. Our experiments show that our ``"muscly-supervised" representation that encodes interaction and attention cues outperforms a visual-only state-of-the-art method MoCo (He et al.,2020), on a variety of target tasks: scene classification (semantic), action recognition (temporal), depth estimation (geometric), dynamics prediction (physics) and walkable surface estimation (affordance). Our code and dataset are available at: https://github.com/ehsanik/muscleTorch.

Contrastive Behavioral Similarity Embeddings for Generalization in Reinforcement Learning

- Rishabh Agarwal, Marlos C. Machado, Pablo Samuel Castro, Marc G Bellemare
- abstract@open-review(Spotlight): Reinforcement learning methods trained on few environments rarely learn policies that generalize to unseen environments. To improve generalization, we incorporate the inherent sequential structure in reinforcement learning into the representation learning process. This approach is orthogonal to recent approaches, which rarely exploit this structure explicitly. Specifically, we introduce a theoretically motivated policy similarity metric (PSM) for measuring behavioral similarity between states. PSM assigns high similarity to states for which the optimal policies in those states as well as in future states are similar. We also present a contrastive representation learning procedure to embed any state similarity metric, which we instantiate with PSM to obtain policy similarity embeddings (PSEs). We demonstrate that PSEs improve generalization on diverse benchmarks, including LQR with spurious correlations, a jumping task from pixels, and Distracting DM Control Suite.

Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images

- Rewon Child
- abstract@open-review(Spotlight): We present a hierarchical VAE that, for the first time, generates samples quickly $\textit{and}$ outperforms the PixelCNN in log-likelihood on all natural image benchmarks. We begin by observing that, in theory, VAEs can actually represent autoregressive models, as well as faster, better models if they exist, when made sufficiently deep. Despite this, autoregressive models have historically outperformed VAEs in log-likelihood. We test if insufficient depth explains why by scaling a VAE to greater stochastic depth than previously explored and evaluating it CIFAR-10, ImageNet, and FFHQ. In comparison to the PixelCNN, these very deep VAEs achieve higher likelihoods, use fewer parameters, generate samples thousands of times faster, and are more easily applied to high-resolution images. Qualitative studies suggest this is because the VAE learns efficient hierarchical visual representations. We release our source code and models at https://github.com/openai/vdvae.

Conservative Safety Critics for Exploration

- Homanga Bharadhwaj, Aviral Kumar, Nicholas Rhinehart, Sergey Levine, Florian Shkurti, Animesh Garg
- abstract@open-review(Poster): Safe exploration presents a major challenge in reinforcement learning (RL): when active data collection requires deploying partially trained policies, we must ensure that these policies avoid catastrophically unsafe regions, while still enabling trial and error learning. In this paper, we target the problem of safe exploration in RL, by learning a conservative safety estimate of environment states through a critic, and provably upper bound the likelihood of catastrophic failures at every training iteration. We theoretically characterize the tradeoff between safety and policy improvement, show that the safety constraints are satisfied with high probability during training, derive provable convergence guarantees for our approach which is no worse asymptotically then standard RL, and empirically demonstrate the efficacy of the proposed approach on a suite of challenging navigation, manipulation, and locomotion tasks. Our results demonstrate that the proposed approach can achieve competitive task performance, while incurring significantly lower catastrophic failure rates during training as compared to prior methods. Videos are at this URL https://sites.google.com/view/conservative-safety-critics/

Multi-Time Attention Networks for Irregularly Sampled Time Series

- Satya Narayan Shukla, Benjamin Marlin
- abstract@open-review(Poster): Irregular sampling occurs in many time series modeling applications where it presents a significant challenge to standard deep learning models. This work is motivated by the analysis of physiological time series data in electronic health records, which are sparse, irregularly sampled, and multivariate. In this paper, we propose a new deep learning framework for this setting that we call Multi-Time Attention Networks. Multi-Time Attention Networks learn an embedding of continuous time values and use an attention mechanism to produce a fixed-length representation of a time series containing a variable number of observations. We investigate the performance of this framework on interpolation and classification tasks using multiple datasets. Our results show that the proposed approach performs as well or better than a range of baseline and recently proposed models while offering significantly faster training times than current state-of-the-art methods.

Graph Traversal with Tensor Functionals: A Meta-Algorithm for Scalable Learning

- Elan Sopher Markowitz, Keshav Balasubramanian, Mehrnoosh Mirtaheri, Sami Abu-El-Haija, Bryan Perozzi, Greg Ver Steeg, Aram Galstyan
- abstract@open-review(Poster): Graph Representation Learning (GRL) methods have impacted fields from chemistry to social science. However, their algorithmic implementations are specialized to specific use-cases e.g. "message passing" methods are run differently from "node embedding" ones. Despite their apparent differences, all these methods utilize the graph structure, and therefore, their learning can be approximated with stochastic graph traversals. We propose Graph Traversal via Tensor Functionals (GTTF), a unifying meta-algorithm framework for easing the implementation of diverse graph algorithms and enabling transparent and efficient scaling to large graphs. GTTF is founded upon a data structure (stored as a sparse tensor) and a stochastic graph traversal algorithm (described using tensor operations). The algorithm is a functional that accept two functions, and can be specialized to obtain a variety of GRL models and objectives, simply by changing those two functions. We show for a wide class of methods, our algorithm learns in an unbiased fashion and, in expectation, approximates the learning as if the specialized implementations were run directly. With these capabilities, we scale otherwise non-scalable methods to set state-of-the-art on large graph datasets while being more efficient than existing GRL libraries -- with only a handful of lines of code for each method specialization.

What they do when in doubt: a study of inductive biases in seq2seq learners

- Eugene Kharitonov, Rahma Chaabouni
- abstract@open-review(Poster): Sequence-to-sequence (seq2seq) learners are widely used, but we still have only limited knowledge about what inductive biases shape the way they generalize. We address that by investigating how popular seq2seq learners generalize in tasks that have high ambiguity in the training data. We use four new tasks to study learners' preferences for memorization, arithmetic, hierarchical, and compositional reasoning. Further, we connect to Solomonoff's theory of induction and propose to use description length as a principled and sensitive measure of inductive biases. In our experimental study, we find that LSTM-based learners can learn to perform counting, addition, and multiplication by a constant from a single training example. Furthermore, Transformer and LSTM-based learners show a bias toward the hierarchical induction over the linear one, while CNN-based learners prefer the opposite. The latter also show a bias toward a compositional generalization over memorization. Finally, across all our experiments, description length proved to be a sensitive measure of inductive biases.

Async-RED: A Provably Convergent Asynchronous Block Parallel Stochastic Method using Deep Denoising Priors

- Yu Sun, Jiaming Liu, Yiran Sun, Brendt Wohlberg, Ulugbek Kamilov
- abstract@open-review(Spotlight): Regularization by denoising (RED) is a recently developed framework for solving inverse problems by integrating advanced denoisers as image priors. Recent work has shown its state-of-the-art performance when combined with pre-trained deep denoisers. However, current RED algorithms are inadequate for parallel processing on multicore systems. We address this issue by proposing a new{asynchronous RED (Async-RED) algorithm that enables asynchronous parallel processing of data, making it significantly faster than its serial counterparts for large-scale inverse problems. The computational complexity of Async-RED is further reduced by using a random subset of measurements at every iteration. We present a complete theoretical analysis of the algorithm by establishing its convergence under explicit assumptions on the data-fidelity and the denoiser. We validate Async-RED on image recovery using pre-trained deep denoisers as priors.

Tomographic Auto-Encoder: Unsupervised Bayesian Recovery of Corrupted Data

- Francesco Tonolini, Pablo Garcia Moreno, Andreas Damianou, Roderick Murray-Smith
- abstract@open-review(Poster): We propose a new probabilistic method for unsupervised recovery of corrupted data. Given a large ensemble of degraded samples, our method recovers accurate posteriors of clean values, allowing the exploration of the manifold of possible reconstructed data and hence characterising the underlying uncertainty. In this set-ting, direct application of classical variational methods often gives rise to collapsed densities that do not adequately explore the solution space. Instead, we derive our novel reduced entropy condition approximate inference method that results in rich posteriors. We test our model in a data recovery task under the common setting of missing values and noise, demonstrating superior performance to existing variational methods for imputation and de-noising with different real data sets. We further show higher classification accuracy after imputation, proving the advantage of propagating uncertainty to downstream tasks with our model.

OPAL: Offline Primitive Discovery for Accelerating Offline Reinforcement Learning

- Anurag Ajay, Aviral Kumar, Pulkit Agrawal, Sergey Levine, Ofir Nachum
- abstract@open-review(Poster): Reinforcement learning (RL) has achieved impressive performance in a variety of online settings in which an agent’s ability to query the environment for transitions and rewards is effectively unlimited. However, in many practical applications, the situation is reversed: an agent may have access to large amounts of undirected offline experience data, while access to the online environment is severely limited. In this work, we focus on this offline setting. Our main insight is that, when presented with offline data composed of a variety of behaviors, an effective way to leverage this data is to extract a continuous space of recurring and temporally extended primitive behaviors before using these primitives for downstream task learning. Primitives extracted in this way serve two purposes: they delineate the behaviors that are supported by the data from those that are not, making them useful for avoiding distributional shift in offline RL; and they provide a degree of temporal abstraction, which reduces the effective horizon yielding better learning in theory, and improved offline RL in practice. In addition to benefiting offline policy optimization, we show that performing offline primitive learning in this way can also be leveraged for improving few-shot imitation learning as well as exploration and transfer in online RL on a variety of benchmark domains. Visualizations and code are available at https://sites.google.com/view/opal-iclr

Knowledge distillation via softmax regression representation learning

- Jing Yang, Brais Martinez, Adrian Bulat, Georgios Tzimiropoulos
- abstract@open-review(Poster): This paper addresses the problem of model compression via knowledge distillation. We advocate for a method that optimizes the output feature of the penultimate layer of the student network and hence is directly related to representation learning. Previous distillation methods which typically impose direct feature matching between the student and the teacher do not take into account the classification problem at hand. On the contrary, our distillation method decouples representation learning and classification and utilizes the teacher's pre-trained classifier to train the student's penultimate layer feature. In particular, for the same input image, we wish the teacher's and student's feature to produce the same output when

passed through the teacher's classifier which is achieved with a simple L_2 loss. Our method is extremely simple to implement and straightforward to train and is shown to consistently outperform previous state-of-the-art methods over a large set of experimental settings including different (a) network architectures, (b) teacher-student capacities, (c) datasets, and (d) domains. The code will be available at \url{https://github.com/jingyang2017/KD_SRRL}.

Large Associative Memory Problem in Neurobiology and Machine Learning

- Dmitry Krotov, John J. Hopfield
- abstract@open-review(Poster): Dense Associative Memories or modern Hopfield networks permit storage and reliable retrieval of an exponentially large (in the dimension of feature space) number of memories. At the same time, their naive implementation is non-biological, since it seemingly requires the existence of many-body synaptic junctions between the neurons. We show that these models are effective descriptions of a more microscopic (written in terms of biological degrees of freedom) theory that has additional (hidden) neurons and only requires two-body interactions between them. For this reason our proposed microscopic theory is a valid model of large associative memory with a degree of biological plausibility. The dynamics of our network and its reduced dimensional equivalent both minimize energy (Lyapunov) functions. When certain dynamical variables (hidden neurons) are integrated out from our microscopic theory, one can recover many of the models that were previously discussed in the literature, e.g. the model presented in "Hopfield Networks is All You Need" paper. We also provide an alternative derivation of the energy function and the update rule proposed in the aforementioned paper and clarify the relationships between various models of this class.

Remembering for the Right Reasons: Explanations Reduce Catastrophic Forgetting

- Sayna Ebrahimi, Suzanne Petryk, Akash Gokul, William Gan, Joseph E. Gonzalez, Marcus Rohrbach, trevor darrell
- abstract@open-review(Poster): The goal of continual learning (CL) is to learn a sequence of tasks without suffering from the phenomenon of catastrophic forgetting. Previous work has shown that leveraging memory in the form of a replay buffer can reduce performance degradation on prior tasks. We hypothesize that forgetting can be further reduced when the model is encouraged to remember the \textit{evidence} for previously made decisions. As a first step towards exploring this hypothesis, we propose a simple novel training paradigm, called Remembering for the Right Reasons (RRR), that additionally stores visual model explanations for each example in the buffer and ensures the model has ``the right reasons" for its predictions by encouraging its explanations to remain consistent with those used to make decisions at training time. Without this constraint, there is a drift in explanations and increase in forgetting as conventional continual learning algorithms learn new tasks. We demonstrate how RRR can be easily added to any memory or regularization-based approach and results in reduced forgetting, and more importantly, improved model explanations. We have evaluated our approach in the standard and few-shot settings and observed a consistent improvement across various CL approaches using different architectures and techniques to generate model explanations and demonstrated our approach showing a promising connection between explainability and continual learning. Our code is available at \url{https://github.com/SaynaEbrahimi/Remembering-for-the-Right-Reasons}.

Deep Networks and the Multiple Manifold Problem

- Sam Buchanan, Dar Gilboa, John Wright
- abstract@open-review(Poster): We study the multiple manifold problem, a binary classification task modeled on applications in machine vision, in which a deep fully-connected neural network is trained to separate two low-dimensional submanifolds of the unit sphere. We provide an analysis of the one-dimensional case, proving for a simple manifold configuration that when the network depth L is large relative to certain geometric and statistical properties of the data, the network width n grows as a sufficiently large polynomial in L, and the number of i.i.d. samples from the manifolds is polynomial in L, randomly-initialized gradient descent rapidly learns to classify the two manifolds perfectly with high probability. Our analysis demonstrates concrete benefits of depth and width in the context of a practically-motivated model problem: the depth acts as a fitting resource, with larger depths corresponding to smoother networks that can more readily separate the class manifolds, and the width acts as a statistical resource, enabling concentration of the randomly-initialized network and its gradients. The argument centers around the "neural tangent kernel" of Jacot et al. and its role in the nonasymptotic analysis of training overparameterized neural networks; to this literature, we contribute essentially optimal rates of concentration for the neural tangent kernel of deep fully-connected ReLU networks, requiring width $n \geq L\,\mathrm{poly}(d_0)$ to achieve uniform concentration of the initial kernel over a d_0-dimensional submanifold of the unit sphere $\mathbb{S}^{n_0-1}$, and a nonasymptotic framework for establishing generalization of networks trained in the "NTK regime" with structured data. The proof makes heavy use of martingale concentration to optimally treat statistical dependencies across layers of the initial random network. This approach should be of use in establishing similar results for other network architectures.

Interpreting Knowledge Graph Relation Representation from Word Embeddings

- Carl Allen, Ivana Balazevic, Timothy Hospedales
- abstract@open-review(Poster): Many models learn representations of knowledge graph data by exploiting its low-rank latent structure, encoding known relations between entities and enabling unknown facts to be inferred. To predict whether a relation holds between entities, embeddings are typically compared in the latent space following a relation-specific mapping. Whilst their predictive performance has steadily improved, how such models capture the underlying latent structure of semantic information remains unexplained. Building on recent theoretical understanding of word embeddings, we categorise knowledge graph relations into three types and for each derive explicit requirements of their representations. We show that empirical properties of relation representations and the relative performance of leading knowledge graph representation methods are justified by our analysis.

Learning perturbation sets for robust machine learning

- Eric Wong, J Zico Kolter
- abstract@open-review(Poster): Although much progress has been made towards robust deep learning, a significant gap in robustness remains between real-world perturbations and more narrowly defined sets typically studied in adversarial defenses. In this paper, we aim to bridge this gap by learning perturbation sets from data, in order to characterize real-world effects for robust training and evaluation. Specifically, we use a conditional generator that defines the perturbation set over a constrained region of the latent space. We formulate desirable properties that measure the quality of a learned perturbation set, and theoretically prove that a conditional variational autoencoder naturally satisfies these criteria. Using this framework, our approach can generate a variety of perturbations at different complexities and scales, ranging from baseline spatial transformations, through common image corruptions, to lighting variations. We measure the quality of our learned perturbation sets both quantitatively and qualitatively, finding that our models are capable of producing a diverse set of meaningful perturbations beyond the limited data seen during training. Finally, we leverage our learned perturbation sets to train models which are empirically and certifiably robust to adversarial image corruptions and adversarial lighting variations, while improving generalization on non-adversarial data. All code and configuration files for reproducing the experiments as well as pretrained model weights can be found at https://github.com/locuslab/perturbation_learning.

Gradient Projection Memory for Continual Learning

- Gobinda Saha, Isha Garg, Kaushik Roy
- abstract@open-review(Oral): The ability to learn continually without forgetting the past tasks is a desired attribute for artificial learning systems. Existing approaches to enable such learning in artificial neural networks usually rely on network growth, importance based weight update or replay of old data from the memory. In contrast, we propose a novel approach where a neural network learns new tasks by taking gradient steps in the orthogonal direction to

the gradient subspaces deemed important for the past tasks. We find the bases of these subspaces by analyzing network representations (activations) after learning each task with Singular Value Decomposition (SVD) in a single shot manner and store them in the memory as Gradient Projection Memory (GPM). With qualitative and quantitative analyses, we show that such orthogonal gradient descent induces minimum to no interference with the past tasks, thereby mitigates forgetting. We evaluate our algorithm on diverse image classification datasets with short and long sequences of tasks and report better or on-par performance compared to the state-of-the-art approaches.

A Trainable Optimal Transport Embedding for Feature Aggregation and its Relationship to Attention

- Grégoire Mialon, Dexiong Chen, Alexandre d'Aspremont, Julien Mairal
- abstract@open-review(Poster): We address the problem of learning on sets of features, motivated by the need of performing pooling operations in long biological sequences of varying sizes, with long-range dependencies, and possibly few labeled data. To address this challenging task, we introduce a parametrized representation of fixed size, which embeds and then aggregates elements from a given input set according to the optimal transport plan between the set and a trainable reference. Our approach scales to large datasets and allows end-to-end training of the reference, while also providing a simple unsupervised learning mechanism with small computational cost. Our aggregation technique admits two useful interpretations: it may be seen as a mechanism related to attention layers in neural networks, or it may be seen as a scalable surrogate of a classical optimal transport-based kernel. We experimentally demonstrate the effectiveness of our approach on biological sequences, achieving state-of-the-art results for protein fold recognition and detection of chromatin profiles tasks, and, as a proof of concept, we show promising results for processing natural language sequences. We provide an open-source implementation of our embedding that can be used alone or as a module in larger learning models at https://github.com/claying/OTK.

RODE: Learning Roles to Decompose Multi-Agent Tasks

- Tonghan Wang, Tarun Gupta, Anuj Mahajan, Bei Peng, Shimon Whiteson, Chongjie Zhang
- abstract@open-review(Poster): Role-based learning holds the promise of achieving scalable multi-agent learning by decomposing complex tasks using roles. However, it is largely unclear how to efficiently discover such a set of roles. To solve this problem, we propose to first decompose joint action spaces into restricted role action spaces by clustering actions according to their effects on the environment and other agents. Learning a role selector based on action effects makes role discovery much easier because it forms a bi-level learning hierarchy: the role selector searches in a smaller role space and at a lower temporal resolution, while role policies learn in significantly reduced primitive action-observation spaces. We further integrate information about action effects into the role policies to boost learning efficiency and policy generalization. By virtue of these advances, our method (1) outperforms the current state-of-the-art MARL algorithms on 9 of the 14 scenarios that comprise the challenging StarCraft II micromanagement benchmark and (2) achieves rapid transfer to new environments with three times the number of agents. Demonstrative videos can be viewed at https://sites.google.com/view/rode-marl.

Scalable Bayesian Inverse Reinforcement Learning

- Alex James Chan, Mihaela van der Schaar
- abstract@open-review(Poster): Bayesian inference over the reward presents an ideal solution to the ill-posed nature of the inverse reinforcement learning problem. Unfortunately current methods generally do not scale well beyond the small tabular setting due to the need for an inner-loop MDP solver, and even non-Bayesian methods that do themselves scale often require extensive interaction with the environment to perform well, being inappropriate for high stakes or costly applications such as healthcare. In this paper we introduce our method, Approximate Variational Reward Imitation Learning (AVRIL), that addresses both of these issues by jointly learning an approximate posterior distribution over the reward that scales to arbitrarily complicated state spaces alongside an appropriate policy in a completely offline manner through a variational approach to said latent reward. Applying our method to real medical data alongside classic control simulations, we demonstrate Bayesian reward inference in environments beyond the scope of current methods, as well as task performance competitive with focused offline imitation learning algorithms.

Learning "What-if" Explanations for Sequential Decision-Making

- Ioana Bica, Daniel Jarrett, Alihan Hüyük, Mihaela van der Schaar
- abstract@open-review(Poster): Building interpretable parameterizations of real-world decision-making on the basis of demonstrated behavior--i.e. trajectories of observations and actions made by an expert maximizing some unknown reward function--is essential for introspecting and auditing policies in different institutions. In this paper, we propose learning explanations of expert decisions by modeling their reward function in terms of preferences with respect to ``"what if" outcomes: Given the current history of observations, what would happen if we took a particular action? To learn these cost-benefit tradeoffs associated with the expert's actions, we integrate counterfactual reasoning into batch inverse reinforcement learning. This offers a principled way of defining reward functions and explaining expert behavior, and also satisfies the constraints of real-world decision-making---where active experimentation is often impossible (e.g. in healthcare). Additionally, by estimating the effects of different actions, counterfactuals readily tackle the off-policy nature of policy evaluation in the batch setting, and can naturally accommodate settings where the expert policies depend on histories of observations rather than just current states. Through illustrative experiments in both real and simulated medical environments, we highlight the effectiveness of our batch, counterfactual inverse reinforcement learning approach in recovering accurate and interpretable descriptions of behavior.

Learning advanced mathematical computations from examples

- Francois Charton, Amaury Hayat, Guillaume Lample
- abstract@open-review(Poster): Using transformers over large generated datasets, we train models to learn mathematical properties of differential systems, such as local stability, behavior at infinity and controllability. We achieve near perfect prediction of qualitative characteristics, and good approximations of numerical features of the system. This demonstrates that neural networks can learn to perform complex computations, grounded in advanced theory, from examples, without built-in mathematical knowledge.

Repurposing Pretrained Models for Robust Out-of-domain Few-Shot Learning

- Namyeong Kwon, Hwidong Na, Gabriel Huang, Simon Lacoste-Julien
- abstract@open-review(Poster): Model-agnostic meta-learning (MAML) is a popular method for few-shot learning but assumes that we have access to the meta-training set. In practice, training on the meta-training set may not always be an option due to data privacy concerns, intellectual property issues, or merely lack of computing resources. In this paper, we consider the novel problem of repurposing pretrained MAML checkpoints to solve new few-shot classification tasks. Because of the potential distribution mismatch, the original MAML steps may no longer be optimal. Therefore we propose an alternative meta-testing procedure and combine MAML gradient steps with adversarial training and uncertainty-based stepsize adaptation. Our method outperforms "vanilla" MAML on same-domain and cross-domains benchmarks using both SGD and Adam optimizers and shows improved robustness to the choice of base stepsize.

Empirical Analysis of Unlabeled Entity Problem in Named Entity Recognition

- Yangming Li, lemao liu, Shuming Shi

- abstract@open-review(Poster): In many scenarios, named entity recognition (NER) models severely suffer from unlabeled entity problem, where the entities of a sentence may not be fully annotated. Through empirical studies performed on synthetic datasets, we find two causes of performance degradation. One is the reduction of annotated entities and the other is treating unlabeled entities as negative instances. The first cause has less impact than the second one and can be mitigated by adopting pretraining language models. The second cause seriously misguides a model in training and greatly affects its performances. Based on the above observations, we propose a general approach, which can almost eliminate the misguidance brought by unlabeled entities. The key idea is to use negative sampling that, to a large extent, avoids training NER models with unlabeled entities. Experiments on synthetic datasets and real-world datasets show that our model is robust to unlabeled entity problem and surpasses prior baselines. On well-annotated datasets, our model is competitive with the state-of-the-art method.

Calibration of Neural Networks using Splines

- Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, Richard Hartley
- abstract@open-review(Poster): Calibrating neural networks is of utmost importance when employing them in safety-critical applications where the downstream decision making depends on the predicted probabilities. Measuring calibration error amounts to comparing two empirical distributions. In this work, we introduce a binning-free calibration measure inspired by the classical Kolmogorov-Smirnov (KS) statistical test in which the main idea is to compare the respective cumulative probability distributions. From this, by approximating the empirical cumulative distribution using a differentiable function via splines, we obtain a recalibration function, which maps the network outputs to actual (calibrated) class assignment probabilities. The spline-fitting is performed using a held-out calibration set and the obtained recalibration function is evaluated on an unseen test set. We tested our method against existing calibration approaches on various image classification datasets and our spline-based recalibration approach consistently outperforms existing methods on KS error as well as other commonly used calibration measures. Code is available online at https://github.com/kartikgupta-at-anu/spline-calibration.

Probing BERT in Hyperbolic Spaces

- Boli Chen, Yao Fu, Guangwei Xu, Pengjun Xie, Chuanqi Tan, Mosha Chen, Liping Jing
- abstract@open-review(Poster): Recently, a variety of probing tasks are proposed to discover linguistic properties learned in contextualized word embeddings. Many of these works implicitly assume these embeddings lay in certain metric spaces, typically the Euclidean space. This work considers a family of geometrically special spaces, the hyperbolic spaces, that exhibit better inductive biases for hierarchical structures and may better reveal linguistic hierarchies encoded in contextualized representations. We introduce a $\textit{Poincaré probe}$, a structural probe projecting these embeddings into a Poincaré subspace with explicitly defined hierarchies. We focus on two probing objectives: (a) dependency trees where the hierarchy is defined as head-dependent structures; (b) lexical sentiments where the hierarchy is defined as the polarity of words (positivity and negativity). We argue that a key desideratum of a probe is its sensitivity to the existence of linguistic structures. We apply our probes on BERT, a typical contextualized embedding model. In a syntactic subspace, our probe better recovers tree structures than Euclidean probes, revealing the possibility that the geometry of BERT syntax may not necessarily be Euclidean. In a sentiment subspace, we reveal two possible meta-embeddings for positive and negative sentiments and show how lexically-controlled contextualization would change the geometric localization of embeddings. We demonstrate the findings with our Poincaré probe via extensive experiments and visualization. Our results can be reproduced at https://github.com/FranxYao/PoincareProbe

Refining Deep Generative Models via Discriminator Gradient Flow

- Abdul Fatir Ansari, Ming Liang Ang, Harold Soh
- abstract@open-review(Poster): Deep generative modeling has seen impressive advances in recent years, to the point where it is now commonplace to see simulated samples (e.g., images) that closely resemble real-world data. However, generation quality is generally inconsistent for any given model and can vary dramatically between samples. We introduce Discriminator Gradient $f$low (DG$f$low), a new technique that improves generated samples via the gradient flow of entropy-regularized f-divergences between the real and the generated data distributions. The gradient flow takes the form of a non-linear Fokker-Plank equation, which can be easily simulated by sampling from the equivalent McKean-Vlasov process. By refining inferior samples, our technique avoids wasteful sample rejection used by previous methods (DRS & MH-GAN). Compared to existing works that focus on specific GAN variants, we show our refinement approach can be applied to GANs with vector-valued critics and even other deep generative models such as VAEs and Normalizing Flows. Empirical results on multiple synthetic, image, and text datasets demonstrate that DG$f$low leads to significant improvement in the quality of generated samples for a variety of generative models, outperforming the state-of-the-art Discriminator Optimal Transport (DOT) and Discriminator Driven Latent Sampling (DDLS) methods.

Coping with Label Shift via Distributionally Robust Optimisation

- Jingzhao Zhang, Aditya Krishna Menon, Andreas Veit, Srinadh Bhojanapalli, Sanjiv Kumar, Suvrit Sra
- abstract@open-review(Poster): The label shift problem refers to the supervised learning setting where the train and test label distributions do not match. Existing work addressing label shift usually assumes access to an unlabelled test sample. This sample may be used to estimate the test label distribution, and to then train a suitably re-weighted classifier. While approaches using this idea have proven effective, their scope is limited as it is not always feasible to access the target domain; further, they require repeated retraining if the model is to be deployed in multiple test environments. Can one instead learn a single classifier that is robust to arbitrary label shifts from a broad family? In this paper, we answer this question by proposing a model that minimises an objective based on distributionally robust optimisation (DRO). We then design and analyse a gradient descent-proximal mirror ascent algorithm tailored for large-scale problems to optimise the proposed objective. Finally, through experiments on CIFAR-100 and ImageNet, we show that our technique can significantly improve performance over a number of baselines in settings where label shift is present.

Variational State-Space Models for Localisation and Dense 3D Mapping in 6 DoF

- Atanas Mirchev, Baris Kayalibay, Patrick van der Smagt, Justin Bayer
- abstract@open-review(Poster): We solve the problem of 6-DoF localisation and 3D dense reconstruction in spatial environments as approximate Bayesian inference in a deep state-space model. Our approach leverages both learning and domain knowledge from multiple-view geometry and rigid-body dynamics. This results in an expressive predictive model of the world, often missing in current state-of-the-art visual SLAM solutions. The combination of variational inference, neural networks and a differentiable raycaster ensures that our model is amenable to end-to-end gradient-based optimisation. We evaluate our approach on realistic unmanned aerial vehicle flight data, nearing the performance of state-of-the-art visual-inertial odometry systems. We demonstrate the applicability of the model to generative prediction and planning.

Few-Shot Bayesian Optimization with Deep Kernel Surrogates

- Martin Wistuba, Josif Grabocka
- abstract@open-review(Poster): Hyperparameter optimization (HPO) is a central pillar in the automation of machine learning solutions and is mainly performed via Bayesian optimization, where a parametric surrogate is learned to approximate the black box response function (e.g. validation error). Unfortunately, evaluating the response function is computationally intensive. As a remedy, earlier work emphasizes the need for transfer learning surrogates which learn to optimize hyperparameters for an algorithm from other tasks. In contrast to previous work, we propose to rethink HPO as a few-shot learning problem in which we train a shared deep surrogate model to quickly adapt (with few response evaluations) to the response function of a new task. We propose the use of a deep kernel network for a Gaussian process surrogate that is meta-learned in an end-to-end fashion in order to jointly

approximate the response functions of a collection of training data sets. As a result, the novel few-shot optimization of our deep kernel surrogate leads to new state-of-the-art results at HPO compared to several recent methods on diverse metadata sets.

i-Mix: A Domain-Agnostic Strategy for Contrastive Representation Learning

- Kibok Lee, Yian Zhu, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin, Honglak Lee
- abstract@open-review(Poster): Contrastive representation learning has shown to be effective to learn representations from unlabeled data. However, much progress has been made in vision domains relying on data augmentations carefully designed using domain knowledge. In this work, we propose i-Mix, a simple yet effective domain-agnostic regularization strategy for improving contrastive representation learning. We cast contrastive learning as training a non-parametric classifier by assigning a unique virtual class to each data in a batch. Then, data instances are mixed in both the input and virtual label spaces, providing more augmented data during training. In experiments, we demonstrate that i-Mix consistently improves the quality of learned representations across domains, including image, speech, and tabular data. Furthermore, we confirm its regularization effect via extensive ablation studies across model and dataset sizes. The code is available at https://github.com/kibok90/imix.

Graph Information Bottleneck for Subgraph Recognition

- Junchi Yu, Tingyang Xu, Yu Rong, Yatao Bian, Junzhou Huang, Ran He
- abstract@open-review(Poster): Given the input graph and its label/property, several key problems of graph learning, such as finding interpretable subgraphs, graph denoising and graph compression, can be attributed to the fundamental problem of recognizing a subgraph of the original one. This subgraph shall be as informative as possible, yet contains less redundant and noisy structure. This problem setting is closely related to the well-known information bottleneck (IB) principle, which, however, has less been studied for the irregular graph data and graph neural networks (GNNs). In this paper, we propose a framework of Graph Information Bottleneck (GIB) for the subgraph recognition problem in deep graph learning. Under this framework, one can recognize the maximally informative yet compressive subgraph, named IB-subgraph. However, the GIB objective is notoriously hard to optimize, mostly due to the intractability of the mutual information of irregular graph data and the unstable optimization process. In order to tackle these challenges, we propose: i) a GIB objective based-on a mutual information estimator for the irregular graph data; ii) a bi-level optimization scheme to maximize the GIB objective; iii) a connectivity loss to stabilize the optimization process. We evaluate the properties of the IB-subgraph in three application scenarios: improvement of graph classification, graph interpretation and graph denoising. Extensive experiments demonstrate that the information-theoretic IB-subgraph enjoys superior graph properties.

Rethinking Positional Encoding in Language Pre-training

- Guolin Ke, Di He, Tie-Yan Liu
- abstract@open-review(Poster): In this work, we investigate the positional encoding methods used in language pre-training (e.g., BERT) and identify several problems in the existing formulations. First, we show that in the absolute positional encoding, the addition operation applied on positional embeddings and word embeddings brings mixed correlations between the two heterogeneous information resources. It may bring unnecessary randomness in the attention and further limit the expressiveness of the model. Second, we question whether treating the position of the symbol \texttt{[CLS]} the same as other words is a reasonable design, considering its special role (the representation of the entire sentence) in the downstream tasks. Motivated from above analysis, we propose a new positional encoding method called \textbf{T}ransformer with \textbf{U}ntied \textbf{P}ositional \textbf{E}ncoding (TUPE). In the self-attention module, TUPE computes the word contextual correlation and positional correlation separately with different parameterizations and then adds them together. This design removes the mixed and noisy correlations over heterogeneous embeddings and offers more expressiveness by using different projection matrices. Furthermore, TUPE unties the \texttt{[CLS]} symbol from other positions, making it easier to capture information from all positions. Extensive experiments and ablation studies on GLUE benchmark demonstrate the effectiveness of the proposed method. Codes and models are released at \url{https://github.com/guolinke/TUPE}.

Practical Massively Parallel Monte-Carlo Tree Search Applied to Molecular Design

- Xiufeng Yang, Tanuj Aasawat, Kazuki Yoshizoe
- abstract@open-review(Poster): It is common practice to use large computational resources to train neural networks, known from many examples, such as reinforcement learning applications. However, while massively parallel computing is often used for training models, it is rarely used to search solutions for combinatorial optimization problems. This paper proposes a novel massively parallel Monte-Carlo Tree Search (MP-MCTS) algorithm that works efficiently for a 1,000 worker scale on a distributed memory environment using multiple compute nodes and applies it to molecular design. This paper is the first work that applies distributed MCTS to a real-world and non-game problem. Existing works on large-scale parallel MCTS show efficient scalability in terms of the number of rollouts up to 100 workers. Still, they suffer from the degradation in the quality of the solutions. MP-MCTS maintains the search quality at a larger scale. By running MP-MCTS on 256 CPU cores for only 10 minutes, we obtained candidate molecules with similar scores to non-parallel MCTS running for 42 hours. Moreover, our results based on parallel MCTS (combined with a simple RNN model) significantly outperform existing state-of-the-art work. Our method is generic and is expected to speed up other applications of MCTS.

Augmenting Physical Models with Deep Networks for Complex Dynamics Forecasting

- Yuan Yin, Vincent LE GUEN, Jérémie DONA, Emmanuel de Bezenac, Ibrahim Ayed, Nicolas THOME, patrick gallinari
- abstract@open-review(Oral): Forecasting complex dynamical phenomena in settings where only partial knowledge of their dynamics is available is a prevalent problem across various scientific fields. While purely data-driven approaches are arguably insufficient in this context, standard physical modeling based approaches tend to be over-simplistic, inducing non-negligible errors. In this work, we introduce the APHYNITY framework, a principled approach for augmenting incomplete physical dynamics described by differential equations with deep data-driven models. It consists in decomposing the dynamics into two components: a physical component accounting for the dynamics for which we have some prior knowledge, and a data-driven component accounting for errors of the physical model. The learning problem is carefully formulated such that the physical model explains as much of the data as possible, while the data-driven component only describes information that cannot be captured by the physical model, no more, no less. This not only provides the existence and uniqueness for this decomposition, but also ensures interpretability and benefits generalization. Experiments made on three important use cases, each representative of a different family of phenomena, i.e. reaction-diffusion equations, wave equations and the non-linear damped pendulum, show that APHYNITY can efficiently leverage approximate physical models to accurately forecast the evolution of the system and correctly identify relevant physical parameters.

When does preconditioning help or hurt generalization?

- Shun-ichi Amari, Jimmy Ba, Roger Baker Grosse, Xuechen Li, Atsushi Nitanda, Taiji Suzuki, Denny Wu, Ji Xu
- abstract@open-review(Poster): While second order optimizers such as natural gradient descent (NGD) often speed up optimization, their effect on generalization has been called into question. This work presents a more nuanced view on how the \textit{implicit bias} of optimizers affects the comparison of generalization properties. We provide an exact asymptotic bias-variance decomposition of the generalization error of preconditioned ridgeless regression in the overparameterized regime, and consider the inverse population Fisher information matrix (used in NGD) as a particular example. We determine the optimal preconditioner $\boldsymbol{P}$ for both the bias and variance, and find that the relative generalization performance of different optimizers depends on label noise and ``shape'' of the signal (true parameters): when the labels are noisy, the model is misspecified, or the signal is misaligned with the features, NGD can achieve lower risk; conversely, GD generalizes better under clean labels, a well-specified model, or

aligned signal. Based on this analysis, we discuss several approaches to manage the bias-variance tradeoff, and the potential benefit of interpolating between first- and second-order updates. We then extend our analysis to regression in the reproducing kernel Hilbert space and demonstrate that preconditioning can lead to more efficient decrease in the population risk. Lastly, we empirically compare the generalization error of first- and second-order optimizers in neural network experiments, and observe robust trends matching our theoretical analysis.

A Good Image Generator Is What You Need for High-Resolution Video Synthesis

- Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N. Metaxas, Sergey Tulyakov
- abstract@open-review(Spotlight): Image and video synthesis are closely related areas aiming at generating content from noise. While rapid progress has been demonstrated in improving image-based models to handle large resolutions, high-quality renderings, and wide variations in image content, achieving comparable video generation results remains problematic. We present a framework that leverages contemporary image generators to render high-resolution videos. We frame the video synthesis problem as discovering a trajectory in the latent space of a pre-trained and fixed image generator. Not only does such a framework render high-resolution videos, but it also is an order of magnitude more computationally efficient. We introduce a motion generator that discovers the desired trajectory, in which content and motion are disentangled. With such a representation, our framework allows for a broad range of applications, including content and motion manipulation. Furthermore, we introduce a new task, which we call cross-domain video synthesis, in which the image and motion generators are trained on disjoint datasets belonging to different domains. This allows for generating moving objects for which the desired video data is not available. Extensive experiments on various datasets demonstrate the advantages of our methods over existing video generation techniques. Code will be released at https://github.com/snap-research/MoCoGAN-HD.

ARMOURED: Adversarially Robust MOdels using Unlabeled data by REgularizing Diversity

- Kangkang Lu, Cuong Manh Nguyen, Xun Xu, Kiran Krishnamachari, Yu Jing Goh, Chuan-Sheng Foo
- abstract@open-review(Poster): Adversarial attacks pose a major challenge for modern deep neural networks. Recent advancements show that adversarially robust generalization requires a large amount of labeled data for training. If annotation becomes a burden, can unlabeled data help bridge the gap? In this paper, we propose ARMOURED, an adversarially robust training method based on semi-supervised learning that consists of two components. The first component applies multi-view learning to simultaneously optimize multiple independent networks and utilizes unlabeled data to enforce labeling consistency. The second component reduces adversarial transferability among the networks via diversity regularizers inspired by determinantal point processes and entropy maximization. Experimental results show that under small perturbation budgets, ARMOURED is robust against strong adaptive adversaries. Notably, ARMOURED does not rely on generating adversarial samples during training. When used in combination with adversarial training, ARMOURED yields competitive performance with the state-of-the-art adversarially-robust benchmarks on SVHN and outperforms them on CIFAR-10, while offering higher clean accuracy.

Learning Energy-Based Generative Models via Coarse-to-Fine Expanding and Sampling

- Yang Zhao, Jianwen Xie, Ping Li
- abstract@open-review(Poster): Energy-based models (EBMs) parameterized by neural networks can be trained by the Markov chain Monte Carlo (MCMC) sampling-based maximum likelihood estimation. Despite the recent significant success of EBMs in image generation, the current approaches to train EBMs are unstable and have difficulty synthesizing diverse and high-fidelity images. In this paper, we propose to train EBMs via a multistage coarse-to-fine expanding and sampling strategy, which starts with learning a coarse-level EBM from images at low resolution and then gradually transits to learn a finer-level EBM from images at higher resolution by expanding the energy function as the learning progresses. The proposed framework is computationally efficient with smooth learning and sampling. It achieves the best performance on image generation amongst all EBMs and is the first successful EBM to synthesize high-fidelity images at 512×512 resolution. It can also be useful for image restoration and out-of-distribution detection. Lastly, the proposed framework is further generalized to the one-sided unsupervised image-to-image translation and beats baseline methods in terms of model size and training budget. We also present a gradient-based generative saliency method to interpret the translation dynamics.

Undistillable: Making A Nasty Teacher That CANNOT teach students

- Haoyu Ma, Tianlong Chen, Ting-Kuei Hu, Chenyu You, Xiaohui Xie, Zhangyang Wang
- abstract@open-review(Spotlight): Knowledge Distillation (KD) is a widely used technique to transfer knowledge from pre-trained teacher models to (usually more lightweight) student models. However, in certain situations, this technique is more of a curse than a blessing. For instance, KD poses a potential risk of exposing intellectual properties (IPs): even if a trained machine learning model is released in ``black boxes'' (e.g., as executable software or APIs without open-sourcing code), it can still be replicated by KD through imitating input-output behaviors. To prevent this unwanted effect of KD, this paper introduces and investigates a concept called $\textit{Nasty Teacher}$: a specially trained teacher network that yields nearly the same performance as a normal one, but would significantly degrade the performance of student models learned by imitating it. We propose a simple yet effective algorithm to build the nasty teacher, called $\textit{self-undermining knowledge distillation}$. Specifically, we aim to maximize the difference between the output of the nasty teacher and a normal pre-trained network. Extensive experiments on several datasets demonstrate that our method is effective on both standard KD and data-free KD, providing the desirable KD-immunity to model owners for the first time. We hope our preliminary study can draw more awareness and interest in this new practical problem of both social and legal importance. Our codes and pre-trained models can be found at: $\url{https://github.com/VITA-Group/Nasty-Teacher}$.

Multi-Prize Lottery Ticket Hypothesis: Finding Accurate Binary Neural Networks by Pruning A Randomly Weighted Network

- James Diffenderfer, Bhavya Kailkhura
- abstract@open-review(Poster): Recently, Frankle & Carbin (2019) demonstrated that randomly-initialized dense networks contain subnetworks that once found can be trained to reach test accuracy comparable to the trained dense network. However, finding these high performing trainable subnetworks is expensive, requiring iterative process of training and pruning weights. In this paper, we propose (and prove) a stronger Multi-Prize Lottery Ticket Hypothesis:

A sufficiently over-parameterized neural network with random weights contains several subnetworks (winning tickets) that (a) have comparable accuracy to a dense target network with learned weights (prize 1), (b) do not require any further training to achieve prize 1 (prize 2), and (c) is robust to extreme forms of quantization (i.e., binary weights and/or activation) (prize 3).

This provides a new paradigm for learning compact yet highly accurate binary neural networks simply by pruning and quantizing randomly weighted full precision neural networks. We also propose an algorithm for finding multi-prize tickets (MPTs) and test it by performing a series of experiments on CIFAR-10 and ImageNet datasets. Empirical results indicate that as models grow deeper and wider, multi-prize tickets start to reach similar (and sometimes even higher) test accuracy compared to their significantly larger and full-precision counterparts that have been weight-trained. Without ever updating the weight values, our MPTs-1/32 not only set new binary weight network state-of-the-art (SOTA) Top-1 accuracy -- 94.8% on CIFAR-10 and 74.03% on ImageNet -- but also outperform their full-precision counterparts by 1.78% and 0.76%, respectively. Further, our MPT-1/1 achieves SOTA Top-1 accuracy (91.9%) for binary neural networks on CIFAR-10. Code and pre-trained models are available at: https://github.com/chrundle/biprop.

Learning Accurate Entropy Model with Global Reference for Image Compression

- Yichen Qian, Zhiyu Tan, Xiuyu Sun, Ming Lin, Dongyang Li, Zhenhong Sun, Li Hao, Rong Jin
- abstract@open-review(Poster): In recent deep image compression neural networks, the entropy model plays a critical role in estimating the prior distribution of deep image encodings. Existing methods combine hyperprior with local context in the entropy estimation function. This greatly limits their performance due to the absence of a global vision. In this work, we propose a novel Global Reference Model for image compression to effectively leverage both the local and the global context information, leading to an enhanced compression rate. The proposed method scans decoded latents and then finds the most relevant latent to assist the distribution estimating of the current latent. A by-product of this work is the innovation of a mean-shifting GDN module that further improves the performance. Experimental results demonstrate that the proposed model outperforms the rate-distortion performance of most of the state-of-the-art methods in the industry.

Convex Potential Flows: Universal Probability Distributions with Optimal Transport and Convex Optimization

- Chin-Wei Huang, Ricky T. Q. Chen, Christos Tsirigotis, Aaron Courville
- abstract@open-review(Poster): Flow-based models are powerful tools for designing probabilistic models with tractable density. This paper introduces Convex Potential Flows (CP-Flow), a natural and efficient parameterization of invertible models inspired by the optimal transport (OT) theory. CP-Flows are the gradient map of a strongly convex neural potential function. The convexity implies invertibility and allows us to resort to convex optimization to solve the convex conjugate for efficient inversion. To enable maximum likelihood training, we derive a new gradient estimator of the log-determinant of the Jacobian, which involves solving an inverse-Hessian vector product using the conjugate gradient method. The gradient estimator has constant-memory cost, and can be made effectively unbiased by reducing the error tolerance level of the convex optimization routine. Theoretically, we prove that CP-Flows are universal density approximators and are optimal in the OT sense. Our empirical results show that CP-Flow performs competitively on standard benchmarks of density estimation and variational inference.

Greedy-GQ with Variance Reduction: Finite-time Analysis and Improved Complexity

- Shaocong Ma, Ziyi Chen, Yi Zhou, Shaofeng Zou
- abstract@open-review(Poster): Greedy-GQ is a value-based reinforcement learning (RL) algorithm for optimal control. Recently, the finite-time analysis of Greedy-GQ has been developed under linear function approximation and Markovian sampling, and the algorithm is shown to achieve an ϵ-stationary point with a sample complexity in the order of $\mathcal{O}(\epsilon^{-3})$. Such a high sample complexity is due to the large variance induced by the Markovian samples. In this paper, we propose a variance-reduced Greedy-GQ (VR-Greedy-GQ) algorithm for off-policy optimal control. In particular, the algorithm applies the SVRG-based variance reduction scheme to reduce the stochastic variance of the two time-scale updates. We study the finite-time convergence of VR-Greedy-GQ under linear function approximation and Markovian sampling and show that the algorithm achieves a much smaller bias and variance error than the original Greedy-GQ. In particular, we prove that VR-Greedy-GQ achieves an improved sample complexity that is in the order of $\mathcal{O}(\epsilon^{-2})$. We further compare the performance of VR-Greedy-GQ with that of Greedy-GQ in various RL experiments to corroborate our theoretical findings.

Large Batch Simulation for Deep Reinforcement Learning

- Brennan Shacklett, Erik Wijmans, Aleksei Petrenko, Manolis Savva, Dhruv Batra, Vladlen Koltun, Kayvon Fatahalian
- abstract@open-review(Poster): We accelerate deep reinforcement learning-based training in visually complex 3D environments by two orders of magnitude over prior work, realizing end-to-end training speeds of over 19,000 frames of experience per second on a single GPU and up to 72,000 frames per second on a single eight-GPU machine. The key idea of our approach is to design a 3D renderer and embodied navigation simulator around the principle of “batch simulation”: accepting and executing large batches of requests simultaneously. Beyond exposing large amounts of work at once, batch simulation allows implementations to amortize in-memory storage of scene assets, rendering work, data loading, and synchronization costs across many simulation requests, dramatically improving the number of simulated agents per GPU and overall simulation throughput. To balance DNN inference and training costs with faster simulation, we also build a computationally efficient policy DNN that maintains high task performance, and modify training algorithms to maintain sample efficiency when training with large mini-batches. By combining batch simulation and DNN performance optimizations, we demonstrate that PointGoal navigation agents can be trained in complex 3D environments on a single GPU in 1.5 days to 97% of the accuracy of agents trained on a prior state-of-the-art system using a 64-GPU cluster over three days. We provide open-source reference implementations of our batch 3D renderer and simulator to facilitate incorporation of these ideas into RL systems.

Hopper: Multi-hop Transformer for Spatiotemporal Reasoning

- Honglu Zhou, Asim Kadav, Farley Lai, Alexandru Niculescu-Mizil, Martin Renqiang Min, Mubbasir Kapadia, Hans Peter Graf
- abstract@open-review(Poster): This paper considers the problem of spatiotemporal object-centric reasoning in videos. Central to our approach is the notion of object permanence, i.e., the ability to reason about the location of objects as they move through the video while being occluded, contained or carried by other objects. Existing deep learning based approaches often suffer from spatiotemporal biases when applied to video reasoning problems. We propose Hopper, which uses a Multi-hop Transformer for reasoning object permanence in videos. Given a video and a localization query, Hopper reasons over image and object tracks to automatically hop over critical frames in an iterative fashion to predict the final position of the object of interest. We demonstrate the effectiveness of using a contrastive loss to reduce spatiotemporal biases. We evaluate over CATER dataset and find that Hopper achieves 73.2% Top-1 accuracy using just 1 FPS by hopping through just a few critical frames. We also demonstrate Hopper can perform long-term reasoning by building a CATER-h dataset that requires multi-step reasoning to localize objects of interest correctly.

Efficient Reinforcement Learning in Factored MDPs with Application to Constrained RL

- Xiaoyu Chen, Jiachen Hu, Lihong Li, Liwei Wang
- abstract@open-review(Poster): Reinforcement learning (RL) in episodic, factored Markov decision processes (FMDPs) is studied. We propose an algorithm called FMDP-BF, which leverages the factorization structure of FMDP. The regret of FMDP-BF is shown to be exponentially smaller than that of optimal algorithms designed for non-factored MDPs, and improves on the best previous result for FMDPs~\citep{osband2014near} by a factor of $\sqrt{nH|\mathcal{S}_i|}$, where $|\mathcal{S}_i|$ is the cardinality of the factored state subspace, H is the planning horizon and n is the number of factored transition. To show the optimality of our bounds, we also provide a lower bound for FMDP, which indicates that our algorithm is near-optimal w.r.t. timestep T, horizon H and factored state-action subspace cardinality. Finally, as an application, we study a new formulation of constrained RL, known as RL with knapsack constraints (RLwK), and provides the first sample-efficient algorithm based on FMDP-BF.

Unbiased Teacher for Semi-Supervised Object Detection

- Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, Peter Vajda
- abstract@open-review(Poster): Semi-supervised learning, i.e., training networks with both labeled and unlabeled data, has made significant progress recently. However, existing works have primarily focused on image classification tasks and neglected object detection which requires more annotation effort. In this work, we revisit the Semi-Supervised Object Detection (SS-OD) and identify the pseudo-labeling bias issue in SS-OD. To address this, we introduce Unbiased Teacher, a simple yet effective approach that jointly trains a student and a gradually progressing teacher in a mutually-beneficial manner. Together with a class-balance loss to downweight overly confident pseudo-labels, Unbiased Teacher consistently improved state-of-the-art methods by significant margins on COCO-standard, COCO-additional, and VOC datasets. Specifically, Unbiased Teacher achieves 6.8 absolute mAP

improvements against state-of-the-art method when using 1% of labeled data on MS-COCO, achieves around 10 mAP improvements against the supervised baseline when using only 0.5, 1, 2% of labeled data on MS-COCO.

Human-Level Performance in No-Press Diplomacy via Equilibrium Search

- Jonathan Gray, Adam Lerer, Anton Bakhtin, Noam Brown
- abstract@open-review(Oral): Prior AI breakthroughs in complex games have focused on either the purely adversarial or purely cooperative settings. In contrast, Diplomacy is a game of shifting alliances that involves both cooperation and competition. For this reason, Diplomacy has proven to be a formidable research challenge. In this paper we describe an agent for the no-press variant of Diplomacy that combines supervised learning on human data with one-step lookahead search via regret minimization. Regret minimization techniques have been behind previous AI successes in adversarial games, most notably poker, but have not previously been shown to be successful in large-scale games involving cooperation. We show that our agent greatly exceeds the performance of past no-press Diplomacy bots, is unexploitable by expert humans, and ranks in the top 2% of human players when playing anonymous games on a popular Diplomacy website.

How Neural Networks Extrapolate: From Feedforward to Graph Neural Networks

- Keyulu Xu, Mozhi Zhang, Jingling Li, Simon Shaolei Du, Ken-Ichi Kawarabayashi, Stefanie Jegelka
- abstract@open-review(Oral): We study how neural networks trained by gradient descent extrapolate, i.e., what they learn outside the support of the training distribution. Previous works report mixed empirical results when extrapolating with neural networks: while feedforward neural networks, a.k.a. multilayer perceptrons (MLPs), do not extrapolate well in certain simple tasks, Graph Neural Networks (GNNs) -- structured networks with MLP modules -- have shown some success in more complex tasks. Working towards a theoretical explanation, we identify conditions under which MLPs and GNNs extrapolate well. First, we quantify the observation that ReLU MLPs quickly converge to linear functions along any direction from the origin, which implies that ReLU MLPs do not extrapolate most nonlinear functions. But, they can provably learn a linear target function when the training distribution is sufficiently diverse. Second, in connection to analyzing the successes and limitations of GNNs, these results suggest a hypothesis for which we provide theoretical and empirical evidence: the success of GNNs in extrapolating algorithmic tasks to new data (e.g., larger graphs or edge weights) relies on encoding task-specific non-linearities in the architecture or features. Our theoretical analysis builds on a connection of over-parameterized networks to the neural tangent kernel. Empirically, our theory holds across different training settings.

MELR: Meta-Learning via Modeling Episode-Level Relationships for Few-Shot Learning

- Nanyi Fei, Zhiwu Lu, Tao Xiang, Songfang Huang
- abstract@open-review(Poster): Most recent few-shot learning (FSL) approaches are based on episodic training whereby each episode samples few training instances (shots) per class to imitate the test condition. However, this strict adhering to test condition has a negative side effect, that is, the trained model is susceptible to the poor sampling of few shots. In this work, for the first time, this problem is addressed by exploiting inter-episode relationships. Specifically, a novel meta-learning via modeling episode-level relationships (MELR) framework is proposed. By sampling two episodes containing the same set of classes for meta-training, MELR is designed to ensure that the meta-learned model is robust against the presence of poorly-sampled shots in the meta-test stage. This is achieved through two key components: (1) a Cross-Episode Attention Module (CEAM) to improve the ability of alleviating the effects of poorly-sampled shots, and (2) a Cross-Episode Consistency Regularization (CECR) to enforce that the two classifiers learned from the two episodes are consistent even when there are unrepresentative instances. Extensive experiments for non-transductive standard FSL on two benchmarks show that our MELR achieves 1.0%-5.0% improvements over the baseline (i.e., ProtoNet) used for FSL in our model and outperforms the latest competitors under the same settings.

Partitioned Learned Bloom Filters

- Kapil Vaidya, Eric Knorr, Michael Mitzenmacher, Tim Kraska
- abstract@open-review(Poster): Bloom filters are space-efficient probabilistic data structures that are used to test whether an element is a member of a set, and may return false positives. Recently, variations referred to as learned Bloom filters were developed that can provide improved performance in terms of the rate of false positives, by using a learned model for the represented set. However, previous methods for learned Bloom filters do not take full advantage of the learned model. Here we show how to frame the problem of optimal model utilization as an optimization problem, and using our framework derive algorithms that can achieve near-optimal performance in many cases.

Wasserstein Embedding for Graph Learning

- Soheil Kolouri, Navid Naderializadeh, Gustavo K. Rohde, Heiko Hoffmann
- abstract@open-review(Poster): We present Wasserstein Embedding for Graph Learning (WEGL), a novel and fast framework for embedding entire graphs in a vector space, in which various machine learning models are applicable for graph-level prediction tasks. We leverage new insights on defining similarity between graphs as a function of the similarity between their node embedding distributions. Specifically, we use the Wasserstein distance to measure the dissimilarity between node embeddings of different graphs. Unlike prior work, we avoid pairwise calculation of distances between graphs and reduce the computational complexity from quadratic to linear in the number of graphs. WEGL calculates Monge maps from a reference distribution to each node embedding and, based on these maps, creates a fixed-sized vector representation of the graph. We evaluate our new graph embedding approach on various benchmark graph-property prediction tasks, showing state-of-the-art classification performance while having superior computational efficiency. The code is available at https://github.com/navid-naderi/WEGL.

High-Capacity Expert Binary Networks

- Adrian Bulat, Brais Martinez, Georgios Tzimiropoulos
- abstract@open-review(Poster): Network binarization is a promising hardware-aware direction for creating efficient deep models. Despite its memory and computational advantages, reducing the accuracy gap between binary models and their real-valued counterparts remains an unsolved challenging research problem. To this end, we make the following 3 contributions: (a) To increase model capacity, we propose Expert Binary Convolution, which, for the first time, tailors conditional computing to binary networks by learning to select one data-specific expert binary filter at a time conditioned on input features. (b) To increase representation capacity, we propose to address the inherent information bottleneck in binary networks by introducing an efficient width expansion mechanism which keeps the binary operations within the same budget. (c) To improve network design, we propose a principled binary network growth mechanism that unveils a set of network topologies of favorable properties. Overall, our method improves upon prior work, with no increase in computational cost, by $\sim6 \%$, reaching a groundbreaking $\sim 71\%$ on ImageNet classification. Code will be made available $\href{https://www.adrianbulat.com/binary-networks}{here}$.

SAFENet: A Secure, Accurate and Fast Neural Network Inference

- Qian Lou, Yilin Shen, Hongxia Jin, Lei Jiang
- abstract@open-review(Poster): The advances in neural networks have driven many companies to provide prediction services to users in a wide range of applications. However, current prediction systems raise privacy concerns regarding the user's private data. A cryptographic neural network inference

service is an efficient way to allow two parties to execute neural network inference without revealing either party's data or model. Nevertheless, existing cryptographic neural network inference services suffer from huge running latency; in particular, the latency of communication-expensive cryptographic activation function is 3 orders of magnitude higher than plaintext-domain activation function. And activations are the necessary components of the modern neural networks. Therefore, slow cryptographic activation has become the primary obstacle of efficient cryptographic inference.

In this paper, we propose a new technique, called SAFENet, to enable a Secure, Accurate and Fast nEural Network inference service. To speedup secure inference and guarantee inference accuracy, SAFENet includes channel-wise activation approximation with multiple-degree options. This is implemented by keeping the most useful activation channels and replacing the remaining, less useful, channels with various-degree polynomials. SAFENet also supports mixed-precision activation approximation by automatically assigning different replacement ratios to various layer; further increasing the approximation ratio and reducing inference latency. Our experimental results show SAFENet obtains the state-of-the-art inference latency and performance, reducing latency by $38\% \sim 61\%$ or improving accuracy by $1.8\% \sim 4\%$ over prior techniques on various encrypted datasets.

Learning Manifold Patch-Based Representations of Man-Made Shapes

- Dmitriy Smirnov, Mikhail Bessmeltsev, Justin Solomon
- abstract@open-review(Poster): Choosing the right representation for geometry is crucial for making 3D models compatible with existing applications. Focusing on piecewise-smooth man-made shapes, we propose a new representation that is usable in conventional CAD modeling pipelines and can also be learned by deep neural networks. We demonstrate its benefits by applying it to the task of sketch-based modeling. Given a raster image, our system infers a set of parametric surfaces that realize the input in 3D. To capture piecewise smooth geometry, we learn a special shape representation: a deformable parametric template composed of Coons patches. Naively training such a system, however, is hampered by non-manifold artifacts in the parametric shapes and by a lack of data. To address this, we introduce loss functions that bias the network to output non-self-intersecting shapes and implement them as part of a fully self-supervised system, automatically generating both shape templates and synthetic training data. We develop a testbed for sketch-based modeling, demonstrate shape interpolation, and provide comparison to related work.

Universal approximation power of deep residual neural networks via nonlinear control theory

- Paulo Tabuada, Bahman Gharesifard
- abstract@open-review(Poster): In this paper, we explain the universal approximation capabilities of deep residual neural networks through geometric nonlinear control. Inspired by recent work establishing links between residual networks and control systems, we provide a general sufficient condition for a residual network to have the power of universal approximation by asking the activation function, or one of its derivatives, to satisfy a quadratic differential equation. Many activation functions used in practice satisfy this assumption, exactly or approximately, and we show this property to be sufficient for an adequately deep neural network with $n+1$ neurons per layer to approximate arbitrarily well, on a compact set and with respect to the supremum norm, any continuous function from $\mathbb{R}^n$ to $\mathbb{R}^n$. We further show this result to hold for very simple architectures for which the weights only need to assume two values. The first key technical contribution consists of relating the universal approximation problem to controllability of an ensemble of control systems corresponding to a residual network and to leverage classical Lie algebraic techniques to characterize controllability. The second technical contribution is to identify monotonicity as the bridge between controllability of finite ensembles and uniform approximability on compact sets.

Learning Neural Event Functions for Ordinary Differential Equations

- Ricky T. Q. Chen, Brandon Amos, Maximilian Nickel
- abstract@open-review(Poster): The existing Neural ODE formulation relies on an explicit knowledge of the termination time. We extend Neural ODEs to implicitly defined termination criteria modeled by neural event functions, which can be chained together and differentiated through. Neural Event ODEs are capable of modeling discrete and instantaneous changes in a continuous-time system, without prior knowledge of when these changes should occur or how many such changes should exist. We test our approach in modeling hybrid discrete- and continuous- systems such as switching dynamical systems and collision in multi-body systems, and we propose simulation-based training of point processes with applications in discrete control.

Neural Spatio-Temporal Point Processes

- Ricky T. Q. Chen, Brandon Amos, Maximilian Nickel
- abstract@open-review(Poster): We propose a new class of parameterizations for spatio-temporal point processes which leverage Neural ODEs as a computational method and enable flexible, high-fidelity models of discrete events that are localized in continuous time and space. Central to our approach is a combination of continuous-time neural networks with two novel neural architectures, \ie, Jump and Attentive Continuous-time Normalizing Flows. This approach allows us to learn complex distributions for both the spatial and temporal domain and to condition non-trivially on the observed event history. We validate our models on data sets from a wide variety of contexts such as seismology, epidemiology, urban mobility, and neuroscience.

Proximal Gradient Descent-Ascent: Variable Convergence under KŁ Geometry

- Ziyi Chen, Yi Zhou, Tengyu Xu, Yingbin Liang
- abstract@open-review(Poster): The gradient descent-ascent (GDA) algorithm has been widely applied to solve minimax optimization problems. In order to achieve convergent policy parameters for minimax optimization, it is important that GDA generates convergent variable sequences rather than convergent sequences of function value or gradient norm. However, the variable convergence of GDA has been proved only under convexity geometries, and it is lack of understanding in general nonconvex minimax optimization. This paper fills such a gap by studying the convergence of a more general proximal-GDA for regularized nonconvex-strongly-concave minimax optimization. Specifically, we show that proximal-GDA admits a novel Lyapunov function, which monotonically decreases in the minimax optimization process and drives the variable sequences to a critical point. By leveraging this Lyapunov function and the KL geometry that parameterizes the local geometries of general nonconvex functions, we formally establish the variable convergence of proximal-GDA to a certain critical point x^*, *i.e.*, $x_t \to x^*$, y_t\to y^(x^)$. Furthermore, over the full spectrum of the KL-parameterized geometry, we show that proximal-GDA achieves different types of convergence rates ranging from sublinear convergence up to finite-step convergence, depending on the geometry associated with the KL parameter. This is the first theoretical result on the variable convergence for nonconvex minimax optimization.

Adaptive Universal Generalized PageRank Graph Neural Network

- Eli Chien, Jianhao Peng, Pan Li, Olgica Milenkovic
- abstract@open-review(Poster): In many important graph data processing applications the acquired information includes both node features and observations of the graph topology. Graph neural networks (GNNs) are designed to exploit both sources of evidence but they do not optimally trade-off their utility and integrate them in a manner that is also universal. Here, universality refers to independence on homophily or heterophily graph assumptions. We address these issues by introducing a new Generalized PageRank (GPR) GNN architecture that adaptively learns the GPR weights so as to jointly optimize node feature and topological information extraction, regardless of the extent to which the node labels are homophilic or heterophilic. Learned GPR weights automatically adjust to the node label pattern, irrelevant on the type of initialization, and thereby guarantee excellent learning performance for label patterns that are usually hard to handle. Furthermore, they allow one to avoid feature over-smoothing, a process which renders feature information nondiscriminative, without requiring the network to be shallow. Our accompanying theoretical analysis of the GPR-GNN method is

facilitated by novel synthetic benchmark datasets generated by the so-called contextual stochastic block model. We also compare the performance of our GNN architecture with that of several state-of-the-art GNNs on the problem of node-classification, using well-known benchmark homophilic and heterophilic datasets. The results demonstrate that GPR-GNN offers significant performance improvement compared to existing techniques on both synthetic and benchmark data.

Open Question Answering over Tables and Text

- Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Yang Wang, William W. Cohen
- abstract@open-review(Poster): In open question answering (QA), the answer to a question is produced by retrieving and then analyzing documents that might contain answers to the question. Most open QA systems have considered only retrieving information from unstructured text. Here we consider for the first time open QA over {\em both} tabular and textual data and present a new large-scale dataset \emph{Open Table-and-Text Question Answering} (OTT-QA) to evaluate performance on this task. Most questions in OTT-QA require multi-hop inference across tabular data and unstructured text, and the evidence required to answer a question can be distributed in different ways over these two types of input, making evidence retrieval challenging---our baseline model using an iterative retriever and BERT-based reader achieves an exact match score less than 10\%. We then propose two novel techniques to address the challenge of retrieving and aggregating evidence for OTT-QA. The first technique is to use ``early fusion'' to group multiple highly relevant tabular and textual units into a fused block, which provides more context for the retriever to search for. The second technique is to use a cross-block reader to model the cross-dependency between multiple retrieved evidence with global-local sparse attention. Combining these two techniques improves the score significantly, to above 27\%.

Rethinking Attention with Performers

- Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, Adrian Weller
- abstract@open-review(Oral): We introduce Performers, Transformer architectures which can estimate regular (softmax) full-rank-attention Transformers with provable accuracy, but using only linear (as opposed to quadratic) space and time complexity, without relying on any priors such as sparsity or low-rankness. To approximate softmax attention-kernels, Performers use a novel Fast Attention Via positive Orthogonal Random features approach (FAVOR+), which may be of independent interest for scalable kernel methods. FAVOR+ can also be used to efficiently model kernelizable attention mechanisms beyond softmax. This representational power is crucial to accurately compare softmax with other kernels for the first time on large-scale tasks, beyond the reach of regular Transformers, and investigate optimal attention-kernels. Performers are linear architectures fully compatible with regular Transformers and with strong theoretical guarantees: unbiased or nearly-unbiased estimation of the attention matrix, uniform convergence and low estimation variance. We tested Performers on a rich set of tasks stretching from pixel-prediction through text models to protein sequence modeling. We demonstrate competitive results with other examined efficient sparse and dense attention methods, showcasing effectiveness of the novel attention-learning paradigm leveraged by Performers.

Text Generation by Learning from Demonstrations

- Richard Yuanzhe Pang, He He
- abstract@open-review(Poster): Current approaches to text generation largely rely on autoregressive models and maximum likelihood estimation. This paradigm leads to (i) diverse but low-quality samples due to mismatched learning objective and evaluation metric (likelihood vs. quality) and (ii) exposure bias due to mismatched history distributions (gold vs. model-generated). To alleviate these problems, we frame text generation as an offline reinforcement learning (RL) problem with expert demonstrations (i.e., the reference), where the goal is to maximize quality given model-generated histories. We propose GOLD (generation by off-policy learning from demonstrations): an easy-to-optimize algorithm that learns from the demonstrations by importance weighting. Intuitively, GOLD upweights confident tokens and downweights unconfident ones in the reference during training, avoiding optimization issues faced by prior RL approaches that rely on online data collection. According to both automatic and human evaluation, models trained by GOLD outperform those trained by MLE and policy gradient on summarization, question generation, and machine translation. Further, our models are less sensitive to decoding algorithms and alleviate exposure bias.

Support-set bottlenecks for video-text representation learning

- Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F. Henriques, Andrea Vedaldi
- abstract@open-review(Spotlight): The dominant paradigm for learning video-text representations – noise contrastive learning – increases the similarity of the representations of pairs of samples that are known to be related, such as text and video from the same sample, and pushes away the representations of all other pairs. We posit that this last behaviour is too strict, enforcing dissimilar representations even for samples that are semantically-related – for example, visually similar videos or ones that share the same depicted action. In this paper, we propose a novel method that alleviates this by leveraging a generative model to naturally push these related samples together: each sample's caption must be reconstructed as a weighted combination of a support set of visual representations. This simple idea ensures that representations are not overly-specialized to individual samples, are reusable across the dataset, and results in representations that explicitly encode semantics shared between samples, unlike noise contrastive learning. Our proposed method outperforms others by a large margin on MSR-VTT, VATEX, ActivityNet, and MSVD for video-to-text and text-to-video retrieval.

Tilted Empirical Risk Minimization

- Tian Li, Ahmad Beirami, Maziar Sanjabi, Virginia Smith
- abstract@open-review(Poster): Empirical risk minimization (ERM) is typically designed to perform well on the average loss, which can result in estimators that are sensitive to outliers, generalize poorly, or treat subgroups unfairly. While many methods aim to address these problems individually, in this work, we explore them through a unified framework---tilted empirical risk minimization (TERM). In particular, we show that it is possible to flexibly tune the impact of individual losses through a straightforward extension to ERM using a hyperparameter called the tilt. We provide several interpretations of the resulting framework: We show that TERM can increase or decrease the influence of outliers, respectively, to enable fairness or robustness; has variance-reduction properties that can benefit generalization; and can be viewed as a smooth approximation to a superquantile method. We develop batch and stochastic first-order optimization methods for solving TERM, and show that the problem can be efficiently solved relative to common alternatives. Finally, we demonstrate that TERM can be used for a multitude of applications, such as enforcing fairness between subgroups, mitigating the effect of outliers, and handling class imbalance. TERM is not only competitive with existing solutions tailored to these individual problems, but can also enable entirely new applications, such as simultaneously addressing outliers and promoting fairness.

Explainable Subgraph Reasoning for Forecasting on Temporal Knowledge Graphs

- Zhen Han, Peng Chen, Yunpu Ma, Volker Tresp
- abstract@open-review(Poster): Modeling time-evolving knowledge graphs (KGs) has recently gained increasing interest. Here, graph representation learning has become the dominant paradigm for link prediction on temporal KGs. However, the embedding-based approaches largely operate in a black-box fashion, lacking the ability to interpret their predictions. This paper provides a link forecasting framework that reasons over query-relevant subgraphs of temporal KGs and jointly models the structural dependencies and the temporal dynamics. Especially, we propose a temporal relational attention mechanism and a novel reverse representation update scheme to guide the extraction of an enclosing subgraph around the query. The subgraph is expanded by an iterative sampling of temporal neighbors and by attention propagation. Our approach provides human-understandable evidence explaining

the forecast. We evaluate our model on four benchmark temporal knowledge graphs for the link forecasting task. While being more explainable, our model obtains a relative improvement of up to 20 $\%$ on Hits@1 compared to the previous best temporal KG forecasting method. We also conduct a survey with 53 respondents, and the results show that the evidence extracted by the model for link forecasting is aligned with human understanding.

Grounded Language Learning Fast and Slow

- Felix Hill, Olivier Tieleman, Tamara von Glehn, Nathaniel Wong, Hamza Merzic, Stephen Clark
- abstract@open-review(Spotlight): Recent work has shown that large text-based neural language models acquire a surprising propensity for one-shot learning. Here, we show that an agent situated in a simulated 3D world, and endowed with a novel dual-coding external memory, can exhibit similar one-shot word learning when trained with conventional RL algorithms. After a single introduction to a novel object via visual perception and language ("This is a dax"), the agent can manipulate the object as instructed ("Put the dax on the bed"), combining short-term, within-episode knowledge of the nonsense word with long-term lexical and motor knowledge. We find that, under certain training conditions and with a particular memory writing mechanism, the agent's one-shot word-object binding generalizes to novel exemplars within the same ShapeNet category, and is effective in settings with unfamiliar numbers of objects. We further show how dual-coding memory can be exploited as a signal for intrinsic motivation, stimulating the agent to seek names for objects that may be useful later. Together, the results demonstrate that deep neural networks can exploit meta-learning, episodic memory and an explicitly multi-modal environment to account for 'fast-mapping', a fundamental pillar of human cognitive development and a potentially transformative capacity for artificial agents.

Bayesian Context Aggregation for Neural Processes

- Michael Volpp, Fabian Flürenbrock, Lukas Grossberger, Christian Daniel, Gerhard Neumann
- abstract@open-review(Poster): Formulating scalable probabilistic regression models with reliable uncertainty estimates has been a long-standing challenge in machine learning research. Recently, casting probabilistic regression as a multi-task learning problem in terms of conditional latent variable (CLV) models such as the Neural Process (NP) has shown promising results. In this paper, we focus on context aggregation, a central component of such architectures, which fuses information from multiple context data points. So far, this aggregation operation has been treated separately from the inference of a latent representation of the target function in CLV models. Our key contribution is to combine these steps into one holistic mechanism by phrasing context aggregation as a Bayesian inference problem. The resulting Bayesian Aggregation (BA) mechanism enables principled handling of task ambiguity, which is key for efficiently processing context information. We demonstrate on a range of challenging experiments that BA consistently improves upon the performance of traditional mean aggregation while remaining computationally efficient and fully compatible with existing NP-based models.

Conformation-Guided Molecular Representation with Hamiltonian Neural Networks

- Ziyao Li, Shuwen Yang, Guojie Song, Lingsheng Cai
- abstract@open-review(Poster): Well-designed molecular representations (fingerprints) are vital to combine medical chemistry and deep learning. Whereas incorporating 3D geometry of molecules (i.e. conformations) in their representations seems beneficial, current 3D algorithms are still in infancy. In this paper, we propose a novel molecular representation algorithm which preserves 3D conformations of molecules with a Molecular Hamiltonian Network (HamNet). In HamNet, implicit positions and momentums of atoms in a molecule interact in the Hamiltonian Engine following the discretized Hamiltonian equations. These implicit coordinations are supervised with real conformations with translation- & rotation-invariant losses, and further used as inputs to the Fingerprint Generator, a message-passing neural network. Experiments show that the Hamiltonian Engine can well preserve molecular conformations, and that the fingerprints generated by HamNet achieve state-of-the-art performances on MoleculeNet, a standard molecular machine learning benchmark.

GAN "Steerability" without optimization

- Nurit Spingarn, Ron Banner, Tomer Michaeli
- abstract@open-review(Spotlight): Recent research has shown remarkable success in revealing "steering" directions in the latent spaces of pre-trained GANs. These directions correspond to semantically meaningful image transformations (e.g., shift, zoom, color manipulations), and have the same interpretable effect across all categories that the GAN can generate. Some methods focus on user-specified transformations, while others discover transformations in an unsupervised manner. However, all existing techniques rely on an optimization procedure to expose those directions, and offer no control over the degree of allowed interaction between different transformations. In this paper, we show that "steering" trajectories can be computed in closed form directly from the generator's weights without any form of training or optimization. This applies to user-prescribed geometric transformations, as well as to unsupervised discovery of more complex effects. Our approach allows determining both linear and nonlinear trajectories, and has many advantages over previous methods. In particular, we can control whether one transformation is allowed to come on the expense of another (e.g., zoom-in with or without allowing translation to keep the object centered). Moreover, we can determine the natural end-point of the trajectory, which corresponds to the largest extent to which a transformation can be applied without incurring degradation. Finally, we show how transferring attributes between images can be achieved without optimization, even across different categories.

Learning with AMIGo: Adversarially Motivated Intrinsic Goals

- Andres Campero, Roberta Raileanu, Heinrich Kuttler, Joshua B. Tenenbaum, Tim Rocktäschel, Edward Grefenstette
- abstract@open-review(Poster): A key challenge for reinforcement learning (RL) consists of learning in environments with sparse extrinsic rewards. In contrast to current RL methods, humans are able to learn new skills with little or no reward by using various forms of intrinsic motivation. We propose AMIGo, a novel agent incorporating -- as form of meta-learning -- a goal-generating teacher that proposes Adversarially Motivated Intrinsic Goals to train a goal-conditioned "student" policy in the absence of (or alongside) environment reward. Specifically, through a simple but effective "constructively adversarial" objective, the teacher learns to propose increasingly challenging -- yet achievable -- goals that allow the student to learn general skills for acting in a new environment, independent of the task to be solved. We show that our method generates a natural curriculum of self-proposed goals which ultimately allows the agent to solve challenging procedurally-generated tasks where other forms of intrinsic motivation and state-of-the-art RL methods fail.

Training with Quantization Noise for Extreme Model Compression

- Pierre Stock, Angela Fan, Benjamin Graham, Edouard Grave, Rémi Gribonval, Herve Jegou, Armand Joulin
- abstract@open-review(Poster): We tackle the problem of producing compact models, maximizing their accuracy for a given model size. A standard solution is to train networks with Quantization Aware Training, where the weights are quantized during training and the gradients approximated with the Straight-Through Estimator. In this paper, we extend this approach to work with extreme compression methods where the approximations introduced by STE are severe. Our proposal is to only quantize a different random subset of weights during each forward, allowing for unbiased gradients to flow through the other weights. Controlling the amount of noise and its form allows for extreme compression rates while maintaining the performance of the original model. As a result we establish new state-of-the-art compromises between accuracy and model size both in natural language processing and image classification. For example, applying our method to state-of-the-art Transformer and ConvNet architectures, we can achieve 82.5% accuracy on MNLI by compressing RoBERTa to 14 MB and 80.0% top-1 accuracy on ImageNet by compressing an EfficientNet-B3 to 3.3 MB.

Interpreting and Boosting Dropout from a Game-Theoretic View

- Hao Zhang, Sen Li, YinChao Ma, Mingjie Li, Yichen Xie, Quanshi Zhang
- abstract@open-review(Poster): This paper aims to understand and improve the utility of the dropout operation from the perspective of game-theoretical interactions. We prove that dropout can suppress the strength of interactions between input variables of deep neural networks (DNNs). The theoretical proof is also verified by various experiments. Furthermore, we find that such interactions were strongly related to the over-fitting problem in deep learning. So, the utility of dropout can be regarded as decreasing interactions to alleviating the significance of over-fitting. Based on this understanding, we propose the interaction loss to further improve the utility of dropout. Experimental results on various DNNs and datasets have shown that the interaction loss can effectively improve the utility of dropout and boost the performance of DNNs.

VTNet: Visual Transformer Network for Object Goal Navigation

- Heming Du, Xin Yu, Liang Zheng
- abstract@open-review(Poster): Object goal navigation aims to steer an agent towards a target object based on observations of the agent. It is of pivotal importance to design effective visual representations of the observed scene in determining navigation actions. In this paper, we introduce a Visual Transformer Network (VTNet) for learning informative visual representation in navigation. VTNet is a highly effective structure that embodies two key properties for visual representations: First, the relationships among all the object instances in a scene are exploited; Second, the spatial locations of objects and image regions are emphasized so that directional navigation signals can be learned. Furthermore, we also develop a pre-training scheme to associate the visual representations with navigation signals, and thus facilitate navigation policy learning. In a nutshell, VTNet embeds object and region features with their location cues as spatial-aware descriptors and then incorporates all the encoded descriptors through attention operations to achieve informative representation for navigation. Given such visual representations, agents are able to explore the correlations between visual observations and navigation actions. For example, an agent would prioritize `turning right'' over`turning left" when the visual representation emphasizes on the right side of activation map. Experiments in the artificial environment AI2-Thor demonstrate that VTNet significantly outperforms state-of-the-art methods in unseen testing environments.

Exemplary Natural Images Explain CNN Activations Better than State-of-the-Art Feature Visualization

- Judy Borowski, Roland Simon Zimmermann, Judith Schepers, Robert Geirhos, Thomas S. A. Wallis, Matthias Bethge, Wieland Brendel
- abstract@open-review(Poster): Feature visualizations such as synthetic maximally activating images are a widely used explanation method to better understand the information processing of convolutional neural networks (CNNs). At the same time, there are concerns that these visualizations might not accurately represent CNNs' inner workings. Here, we measure how much extremely activating images help humans to predict CNN activations. Using a well-controlled psychophysical paradigm, we compare the informativeness of synthetic images by Olah et al. (2017) with a simple baseline visualization, namely exemplary natural images that also strongly activate a specific feature map. Given either synthetic or natural reference images, human participants choose which of two query images leads to strong positive activation. The experiment is designed to maximize participants' performance, and is the first to probe intermediate instead of final layer representations. We find that synthetic images indeed provide helpful information about feature map activations ($82\pm4\%$ accuracy; chance would be 50%). However, natural images --- originally intended to be a baseline --- outperform these synthetic images by a wide margin ($92\pm2\%$). Additionally, participants are faster and more confident for natural images, whereas subjective impressions about the interpretability of the feature visualizations by Olah et al. (2017) are mixed. The higher informativeness of natural images holds across most layers, for both expert and lay participants as well as for hand- and randomly-picked feature visualizations. Even if only a single reference image is given, synthetic images provide less information than natural images ($65\pm5\%$ vs. $73\pm4\%$). In summary, synthetic images from a popular feature visualization method are significantly less informative for assessing CNN activations than natural images. We argue that visualization methods should improve over this simple baseline.

Multi-Class Uncertainty Calibration via Mutual Information Maximization-based Binning

- Kanil Patel, William H. Beluch, Bin Yang, Michael Pfeiffer, Dan Zhang
- abstract@open-review(Poster): Post-hoc multi-class calibration is a common approach for providing high-quality confidence estimates of deep neural network predictions. Recent work has shown that widely used scaling methods underestimate their calibration error, while alternative Histogram Binning (HB) methods often fail to preserve classification accuracy. When classes have small prior probabilities, HB also faces the issue of severe sample-inefficiency after the conversion into K one-vs-rest class-wise calibration problems. The goal of this paper is to resolve the identified issues of HB in order to provide calibrated confidence estimates using only a small holdout calibration dataset for bin optimization while preserving multi-class ranking accuracy. From an information-theoretic perspective, we derive the I-Max concept for binning, which maximizes the mutual information between labels and quantized logits. This concept mitigates potential loss in ranking performance due to lossy quantization, and by disentangling the optimization of bin edges and representatives allows simultaneous improvement of ranking and calibration performance. To improve the sample efficiency and estimates from a small calibration set, we propose a shared class-wise (sCW) calibration strategy, sharing one calibrator among similar classes (e.g., with similar class priors) so that the training sets of their class-wise calibration problems can be merged to train the single calibrator. The combination of sCW and I-Max binning outperforms the state of the art calibration methods on various evaluation metrics across different benchmark datasets and models, using a small calibration set (e.g., 1k samples for ImageNet).

A Discriminative Gaussian Mixture Model with Sparsity

- Hideaki Hayashi, Seiichi Uchida
- abstract@open-review(Poster): In probabilistic classification, a discriminative model based on the softmax function has a potential limitation in that it assumes unimodality for each class in the feature space. The mixture model can address this issue, although it leads to an increase in the number of parameters. We propose a sparse classifier based on a discriminative GMM, referred to as a sparse discriminative Gaussian mixture (SDGM). In the SDGM, a GMM-based discriminative model is trained via sparse Bayesian learning. Using this sparse learning framework, we can simultaneously remove redundant Gaussian components and reduce the number of parameters used in the remaining components during learning; this learning method reduces the model complexity, thereby improving the generalization capability. Furthermore, the SDGM can be embedded into neural networks (NNs), such as convolutional NNs, and can be trained in an end-to-end manner. Experimental results demonstrated that the proposed method outperformed the existing softmax-based discriminative models.

Trusted Multi-View Classification

- Zongbo Han, Changqing Zhang, Huazhu Fu, Joey Tianyi Zhou
- abstract@open-review(Poster): Multi-view classification (MVC) generally focuses on improving classification accuracy by using information from different views, typically integrating them into a unified comprehensive representation for downstream tasks. However, it is also crucial to dynamically assess the quality of a view for different samples in order to provide reliable uncertainty estimations, which indicate whether predictions can be trusted. To this end, we propose a novel multi-view classification method, termed trusted multi-view classification, which provides a new paradigm for multi-view learning by dynamically integrating different views at an evidence level. The algorithm jointly utilizes multiple views to promote both classification reliability (uncertainty estimation during testing) and robustness (out-of-distribution-awareness during training) by integrating evidence from each view. To achieve this, the Dirichlet distribution is used to model the distribution of the class probabilities, parameterized with evidence from different views and integrated with the Dempster-Shafer theory. The unified learning framework induces accurate uncertainty and accordingly endows the model with both

reliability and robustness for out-of-distribution samples. Extensive experimental results validate the effectiveness of the proposed model in accuracy, reliability and robustness.

IEPT: Instance-Level and Episode-Level Pretext Tasks for Few-Shot Learning

- Manli Zhang, Jianhong Zhang, Zhiwu Lu, Tao Xiang, Mingyu Ding, Songfang Huang
- abstract@open-review(Poster): The need of collecting large quantities of labeled training data for each new task has limited the usefulness of deep neural networks. Given data from a set of source tasks, this limitation can be overcome using two transfer learning approaches: few-shot learning (FSL) and self-supervised learning (SSL). The former aims to learn `how to learn' by designing learning episodes using source tasks to simulate the challenge of solving the target new task with few labeled samples. In contrast, the latter exploits an annotation-free pretext task across all source tasks in order to learn generalizable feature representations. In this work, we propose a novel Instance-level and Episode-level Pretext Task (IEPT) framework that seamlessly integrates SSL into FSL. Specifically, given an FSL episode, we first apply geometric transformations to each instance to generate extended episodes. At the instance-level, transformation recognition is performed as per standard SSL. Importantly, at the episode-level, two SSL-FSL hybrid learning objectives are devised: (1) The consistency across the predictions of an FSL classifier from different extended episodes is maximized as an episode-level pretext task. (2) The features extracted from each instance across different episodes are integrated to construct a single FSL classifier for meta-learning. Extensive experiments show that our proposed model (i.e., FSL with IEPT) achieves the new state-of-the-art.

What Matters for On-Policy Deep Actor-Critic Methods? A Large-Scale Study

- Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphaël Marinier, Leonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, Sylvain Gelly, Olivier Bachem
- abstract@open-review(Oral): In recent years, reinforcement learning (RL) has been successfully applied to many different continuous control tasks. While RL algorithms are often conceptually simple, their state-of-the-art implementations take numerous low- and high-level design decisions that strongly affect the performance of the resulting agents. Those choices are usually not extensively discussed in the literature, leading to discrepancy between published descriptions of algorithms and their implementations. This makes it hard to attribute progress in RL and slows down overall progress [Engstrom'20]. As a step towards filling that gap, we implement >50 such ``"choices" in a unified on-policy deep actor-critic framework, allowing us to investigate their impact in a large-scale empirical study. We train over 250'000 agents in five continuous control environments of different complexity and provide insights and practical recommendations for the training of on-policy deep actor-critic RL agents.

Deep Encoder, Shallow Decoder: Reevaluating Non-autoregressive Machine Translation

- Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, Noah Smith
- abstract@open-review(Poster): Much recent effort has been invested in non-autoregressive neural machine translation, which appears to be an efficient alternative to state-of-the-art autoregressive machine translation on modern GPUs. In contrast to the latter, where generation is sequential, the former allows generation to be parallelized across target token positions. Some of the latest non-autoregressive models have achieved impressive translation quality-speed tradeoffs compared to autoregressive baselines. In this work, we reexamine this tradeoff and argue that autoregressive baselines can be substantially sped up without loss in accuracy. Specifically, we study autoregressive models with encoders and decoders of varied depths. Our extensive experiments show that given a sufficiently deep encoder, a single-layer autoregressive decoder can substantially outperform strong non-autoregressive models with comparable inference speed. We show that the speed disadvantage for autoregressive baselines compared to non-autoregressive methods has been overestimated in three aspects: suboptimal layer allocation, insufficient speed measurement, and lack of knowledge distillation. Our results establish a new protocol for future research toward fast, accurate machine translation. Our code is available at https://github.com/jungokasai/deep-shallow.

Effective Abstract Reasoning with Dual-Contrast Network

- Tao Zhuo, Mohan Kankanhalli
- abstract@open-review(Poster): As a step towards improving the abstract reasoning capability of machines, we aim to solve Raven’s Progressive Matrices (RPM) with neural networks, since solving RPM puzzles is highly correlated with human intelligence. Unlike previous methods that use auxiliary annotations or assume hidden rules to produce appropriate feature representation, we only use the ground truth answer of each question for model learning, aiming for an intelligent agent to have a strong learning capability with a small amount of supervision. Based on the RPM problem formulation, the correct answer filled into the missing entry of the third row/column has to best satisfy the same rules shared between the first two rows/columns.Thus we design a simple yet effective Dual-Contrast Network (DCNet) to exploit the inherent structure of RPM puzzles. Specifically, a rule contrast module is designed to compare the latent rules between the filled row/column and the first two rows/columns; a choice contrast module is designed to increase the relative differences between candidate choices. Experimental results on the RAVEN and PGM datasets show that DCNet outperforms the state-of-the-art methods by a large margin of 5.77%. Further experiments on few training samples and model generalization also show the effectiveness of DCNet. Code is available at https://github.com/visiontao/dcnet.

Noise against noise: stochastic label noise helps combat inherent label noise

- Pengfei Chen, Guangyong Chen, Junjie Ye, jingwei zhao, Pheng-Ann Heng
- abstract@open-review(Spotlight): The noise in stochastic gradient descent (SGD) provides a crucial implicit regularization effect, previously studied in optimization by analyzing the dynamics of parameter updates. In this paper, we are interested in learning with noisy labels, where we have a collection of samples with potential mislabeling. We show that a previously rarely discussed SGD noise, induced by stochastic label noise (SLN), mitigates the effects of inherent label noise. In contrast, the common SGD noise directly applied to model parameters does not. We formalize the differences and connections of SGD noise variants, showing that SLN induces SGD noise dependent on the sharpness of output landscape and the confidence of output probability, which may help escape from sharp minima and prevent overconfidence. SLN not only improves generalization in its simplest form but also boosts popular robust training methods, including sample selection and label correction. Specifically, we present an enhanced algorithm by applying SLN to label correction. Our code is released.

VAEBM: A Symbiosis between Variational Autoencoders and Energy-based Models

- Zhisheng Xiao, Karsten Kreis, Jan Kautz, Arash Vahdat
- abstract@open-review(Spotlight): Energy-based models (EBMs) have recently been successful in representing complex distributions of small images. However, sampling from them requires expensive Markov chain Monte Carlo (MCMC) iterations that mix slowly in high dimensional pixel space. Unlike EBMs, variational autoencoders (VAEs) generate samples quickly and are equipped with a latent space that enables fast traversal of the data manifold. However, VAEs tend to assign high probability density to regions in data space outside the actual data distribution and often fail at generating sharp images. In this paper, we propose VAEBM, a symbiotic composition of a VAE and an EBM that offers the best of both worlds. VAEBM captures the overall mode structure of the data distribution using a state-of-the-art VAE and it relies on its EBM component to explicitly exclude non-data-like regions from the model and refine the image samples. Moreover, the VAE component in VAEBM allows us to speed up MCMC updates by reparameterizing them in the VAE's latent space. Our experimental results show that VAEBM outperforms state-of-the-art VAEs and EBMs in generative quality on several benchmark image datasets by a large margin. It can generate high-quality images as large as 256$\times$256 pixels with short MCMC chains. We also demonstrate that VAEBM provides complete mode coverage and performs well in out-of-distribution detection.

On Position Embeddings in BERT

- Benyou Wang, Lifeng Shang, Christina Lioma, Xin Jiang, Hao Yang, Qun Liu, Jakob Grue Simonsen
- abstract@open-review(Poster): Various Position Embeddings (PEs) have been proposed in Transformer based architectures~(e.g. BERT) to model word order. These are empirically-driven and perform well, but no formal framework exists to systematically study them. To address this, we present three properties of PEs that capture word distance in vector space: translation invariance, monotonicity, and symmetry. These properties formally capture the behaviour of PEs and allow us to reinterpret sinusoidal PEs in a principled way. Moreover, we propose a new probing test (called `identical word probing') and mathematical indicators to quantitatively detect the general attention patterns with respect to the above properties. An empirical evaluation of seven PEs (and their combinations) for classification (GLUE) and span prediction (SQuAD) shows that: (1) both classification and span prediction benefit from translation invariance and local monotonicity, while symmetry slightly decreases performance; (2) The fully-learnable absolute PE performs better in classification, while relative PEs perform better in span prediction. We contribute the first formal and quantitative analysis of desiderata for PEs, and a principled discussion about their correlation to the performance of typical downstream tasks.

Neural Pruning via Growing Regularization

- Huan Wang, Can Qin, Yulun Zhang, Yun Fu
- abstract@open-review(Poster): Regularization has long been utilized to learn sparsity in deep neural network pruning. However, its role is mainly explored in the small penalty strength regime. In this work, we extend its application to a new scenario where the regularization grows large gradually to tackle two central problems of pruning: pruning schedule and weight importance scoring. (1) The former topic is newly brought up in this work, which we find critical to the pruning performance while receives little research attention. Specifically, we propose an L2 regularization variant with rising penalty factors and show it can bring significant accuracy gains compared with its one-shot counterpart, even when the same weights are removed. (2) The growing penalty scheme also brings us an approach to exploit the Hessian information for more accurate pruning without knowing their specific values, thus not bothered by the common Hessian approximation problems. Empirically, the proposed algorithms are easy to implement and scalable to large datasets and networks in both structured and unstructured pruning. Their effectiveness is demonstrated with modern deep neural networks on the CIFAR and ImageNet datasets, achieving competitive results compared to many state-of-the-art algorithms. Our code and trained models are publicly available at https://github.com/mingsun-tse/regularization-pruning.

Mixed-Features Vectors and Subspace Splitting

- Alejandro Pimentel-Alarcón, Daniel L. Pimentel-Alarcón
- abstract@open-review(Poster): Motivated by metagenomics, recommender systems, dictionary learning, and related problems, this paper introduces subspace splitting(SS): the task of clustering the entries of what we call amixed-features vector, that is, a vector whose subsets of coordinates agree with a collection of subspaces. We derive precise identifiability conditions under which SS is well-posed, thus providing the first fundamental theory for this problem. We also propose the first three practical SS algorithms, each with advantages and disadvantages: a random sampling method , a projection-based greedy heuristic , and an alternating Lloyd-type algorithm ; all allow noise, outliers, and missing data. Our extensive experiments outline the performance of our algorithms, and in lack of other SS algorithms, for reference we compare against methods for tightly related problems, like robust matched subspace detection and maximum feasible subsystem, which are special simpler cases of SS.

Hierarchical Reinforcement Learning by Discovering Intrinsic Options

- Jesse Zhang, Haonan Yu, Wei Xu
- abstract@open-review(Poster): We propose a hierarchical reinforcement learning method, HIDIO, that can learn task-agnostic options in a self-supervised manner while jointly learning to utilize them to solve sparse-reward tasks. Unlike current hierarchical RL approaches that tend to formulate goal-reaching low-level tasks or pre-define ad hoc lower-level policies, HIDIO encourages lower-level option learning that is independent of the task at hand, requiring few assumptions or little knowledge about the task structure. These options are learned through an intrinsic entropy minimization objective conditioned on the option sub-trajectories. The learned options are diverse and task-agnostic. In experiments on sparse-reward robotic manipulation and navigation tasks, HIDIO achieves higher success rates with greater sample efficiency than regular RL baselines and two state-of-the-art hierarchical RL methods. Code at: https://github.com/jesbu1/hidio.

Sharper Generalization Bounds for Learning with Gradient-dominated Objective Functions

- Yunwen Lei, Yiming Ying
- abstract@open-review(Poster): Stochastic optimization has become the workhorse behind many successful machine learning applications, which motivates a lot of theoretical analysis to understand its empirical behavior. As a comparison, there is far less work to study the generalization behavior especially in a non-convex learning setting. In this paper, we study the generalization behavior of stochastic optimization by leveraging the algorithmic stability for learning with β-gradient-dominated objective functions. We develop generalization bounds of the order $O(1/(n\beta))$ plus the convergence rate of the optimization algorithm, where n is the sample size. Our stability analysis significantly improves the existing non-convex analysis by removing the bounded gradient assumption and implying better generalization bounds. We achieve this improvement by exploiting the smoothness of loss functions instead of the Lipschitz condition in Charles & Papailiopoulos (2018). We apply our general results to various stochastic optimization algorithms, which show clearly how the variance-reduction techniques improve not only training but also generalization. Furthermore, our discussion explains how interpolation helps generalization for highly expressive models.

Representation Learning for Sequence Data with Deep Autoencoding Predictive Components

- Junwen Bai, Weiran Wang, Yingbo Zhou, Caiming Xiong
- abstract@open-review(Poster): We propose Deep Autoencoding Predictive Components (DAPC) -- a self-supervised representation learning method for sequence data, based on the intuition that useful representations of sequence data should exhibit a simple structure in the latent space. We encourage this latent structure by maximizing an estimate of \emph{predictive information} of latent feature sequences, which is the mutual information between the past and future windows at each time step. In contrast to the mutual information lower bound commonly used by contrastive learning, the estimate of predictive information we adopt is exact under a Gaussian assumption. Additionally, it can be computed without negative sampling. To reduce the degeneracy of the latent space extracted by powerful encoders and keep useful information from the inputs, we regularize predictive information learning with a challenging masked reconstruction loss. We demonstrate that our method recovers the latent space of noisy dynamical systems, extracts predictive features for forecasting tasks, and improves automatic speech recognition when used to pretrain the encoder on large amounts of unlabeled data.

Average-case Acceleration for Bilinear Games and Normal Matrices

- Carles Domingo-Enrich, Fabian Pedregosa, Damien Scieur
- abstract@open-review(Poster): Advances in generative modeling and adversarial learning have given rise to renewed interest in smooth games. However, the absence of symmetry in the matrix of second derivatives poses challenges that are not present in the classical minimization framework. While a rich theory of average-case analysis has been developed for minimization problems, little is known in the context of smooth games. In this work we take a first step towards closing this gap by developing average-case optimal first-order methods for a subset of smooth games. We make the following three main

contributions. First, we show that for zero-sum bilinear games the average-case optimal method is the optimal method for the minimization of the Hamiltonian. Second, we provide an explicit expression for the optimal method corresponding to normal matrices, potentially non-symmetric. Finally, we specialize it to matrices with eigenvalues located in a disk and show a provable speed-up compared to worst-case optimal algorithms. We illustrate our findings through benchmarks with a varying degree of mismatch with our assumptions.

Graph-Based Continual Learning

- Binh Tang, David S. Matteson
- abstract@open-review(Spotlight): Despite significant advances, continual learning models still suffer from catastrophic forgetting when exposed to incrementally available data from non-stationary distributions. Rehearsal approaches alleviate the problem by maintaining and replaying a small episodic memory of previous samples, often implemented as an array of independent memory slots. In this work, we propose to augment such an array with a learnable random graph that captures pairwise similarities between its samples, and use it not only to learn new tasks but also to guard against forgetting. Empirical results on several benchmark datasets show that our model consistently outperforms recently proposed baselines for task-free continual learning.

Learning Task-General Representations with Generative Neuro-Symbolic Modeling

- Reuben Feinman, Brenden M. Lake
- abstract@open-review(Poster): People can learn rich, general-purpose conceptual representations from only raw perceptual inputs. Current machine learning approaches fall well short of these human standards, although different modeling traditions often have complementary strengths. Symbolic models can capture the compositional and causal knowledge that enables flexible generalization, but they struggle to learn from raw inputs, relying on strong abstractions and simplifying assumptions. Neural network models can learn directly from raw data, but they struggle to capture compositional and causal structure and typically must retrain to tackle new tasks. We bring together these two traditions to learn generative models of concepts that capture rich compositional and causal structure, while learning from raw data. We develop a generative neuro-symbolic (GNS) model of handwritten character concepts that uses the control flow of a probabilistic program, coupled with symbolic stroke primitives and a symbolic image renderer, to represent the causal and compositional processes by which characters are formed. The distributions of parts (strokes), and correlations between parts, are modeled with neural network subroutines, allowing the model to learn directly from raw data and express nonparametric statistical relationships. We apply our model to the Omniglot challenge of human-level concept learning, using a background set of alphabets to learn an expressive prior distribution over character drawings. In a subsequent evaluation, our GNS model uses probabilistic inference to learn rich conceptual representations from a single training image that generalize to 4 unique tasks, succeeding where previous work has fallen short.

Is Label Smoothing Truly Incompatible with Knowledge Distillation: An Empirical Study

- Zhiqiang Shen, Zechun Liu, Dejia Xu, Zitian Chen, Kwang-Ting Cheng, Marios Savvides
- abstract@open-review(Poster): This work aims to empirically clarify a recently discovered perspective that label smoothing is incompatible with knowledge distillation. We begin by introducing the motivation behind on how this incompatibility is raised, i.e., label smoothing erases relative information between teacher logits. We provide a novel connection on how label smoothing affects distributions of semantically similar and dissimilar classes. Then we propose a metric to quantitatively measure the degree of erased information in sample's representation. After that, we study its one-sidedness and imperfection of the incompatibility view through massive analyses, visualizations and comprehensive experiments on Image Classification, Binary Networks, and Neural Machine Translation. Finally, we broadly discuss several circumstances wherein label smoothing will indeed lose its effectiveness.

Sparse Quantized Spectral Clustering

- Zhenyu Liao, Romain Couillet, Michael W. Mahoney
- abstract@open-review(Spotlight): Given a large data matrix, sparsifying, quantizing, and/or performing other entry-wise nonlinear operations can have numerous benefits, ranging from speeding up iterative algorithms for core numerical linear algebra problems to providing nonlinear filters to design state-of-the-art neural network models. Here, we exploit tools from random matrix theory to make precise statements about how the eigenspectrum of a matrix changes under such nonlinear transformations. In particular, we show that very little change occurs in the informative eigenstructure, even under drastic sparsification/quantization, and consequently that very little downstream performance loss occurs when working with very aggressively sparsified or quantized spectral clustering problems. We illustrate how these results depend on the nonlinearity, we characterize a phase transition beyond which spectral clustering becomes possible, and we show when such nonlinear transformations can introduce spurious non-informative eigenvectors.

The Unreasonable Effectiveness of Patches in Deep Convolutional Kernels Methods

- Louis THIRY, Michael Arbel, Eugene Belilovsky, Edouard Oyallon
- abstract@open-review(Poster): A recent line of work showed that various forms of convolutional kernel methods can be competitive with standard supervised deep convolutional networks on datasets like CIFAR-10, obtaining accuracies in the range of 87-90% while being more amenable to theoretical analysis. In this work, we highlight the importance of a data-dependent feature extraction step that is key to the obtain good performance in convolutional kernel methods. This step typically corresponds to a whitened dictionary of patches, and gives rise to a data-driven convolutional kernel methods.We extensively study its effect, demonstrating it is the key ingredient for high performance of these methods. Specifically, we show that one of the simplest instances of such kernel methods, based on a single layer of image patches followed by a linear classifier is already obtaining classification accuracies on CIFAR-10 in the same range as previous more sophisticated convolutional kernel methods. We scale this method to the challenging ImageNet dataset, showing such a simple approach can exceed all existing non-learned representation methods. This is a new baseline for object recognition without representation learning methods, that initiates the investigation of convolutional kernel models on ImageNet. We conduct experiments to analyze the dictionary that we used, our ablations showing they exhibit low-dimensional properties.

Graph Coarsening with Neural Networks

- Chen Cai, Dingkang Wang, Yusu Wang
- abstract@open-review(Poster): As large scale-graphs become increasingly more prevalent, it poses significant computational challenges to process, extract and analyze large graph data. Graph coarsening is one popular technique to reduce the size of a graph while maintaining essential properties. Despite rich graph coarsening literature, there is only limited exploration of data-driven method in the field. In this work, we leverage the recent progress of deep learning on graphs for graph coarsening. We first propose a framework for measuring the quality of coarsening algorithm and show that depending on the goal, we need to carefully choose the Laplace operator on the coarse graph and associated projection/lift operators. Motivated by the observation that the current choice of edge weight for the coarse graph may be sub-optimal, we parametrize the weight assignment map with graph neural networks and train it to improve the coarsening quality in an unsupervised way. Through extensive experiments on both synthetic and real networks, we demonstrate that our method significantly improves common graph coarsening methods under various metrics, reduction ratios, graph sizes, and graph types. It generalizes to graphs of larger size (more than $25\times$ of training graphs), adaptive to different losses (both differentiable and non-differentiable), and scales to much larger graphs than previous work.

On the Universality of the Double Descent Peak in Ridgeless Regression

- David Holzmüller
- abstract@open-review(Poster): We prove a non-asymptotic distribution-independent lower bound for the expected mean squared generalization error caused by label noise in ridgeless linear regression. Our lower bound generalizes a similar known result to the overparameterized (interpolating) regime. In contrast to most previous works, our analysis applies to a broad class of input distributions with almost surely full-rank feature matrices, which allows us to cover various types of deterministic or random feature maps. Our lower bound is asymptotically sharp and implies that in the presence of label noise, ridgeless linear regression does not perform well around the interpolation threshold for any of these feature maps. We analyze the imposed assumptions in detail and provide a theory for analytic (random) feature maps. Using this theory, we can show that our assumptions are satisfied for input distributions with a (Lebesgue) density and feature maps given by random deep neural networks with analytic activation functions like sigmoid, tanh, softplus or GELU. As further examples, we show that feature maps from random Fourier features and polynomial kernels also satisfy our assumptions. We complement our theory with further experimental and analytic results.

Deep Repulsive Clustering of Ordered Data Based on Order-Identity Decomposition

- Seon-Ho Lee, Chang-Su Kim
- abstract@open-review(Poster): We propose the deep repulsive clustering (DRC) algorithm of ordered data for effective order learning. First, we develop the order-identity decomposition (ORID) network to divide the information of an object instance into an order-related feature and an identity feature. Then, we group object instances into clusters according to their identity features using a repulsive term. Moreover, we estimate the rank of a test instance, by comparing it with references within the same cluster. Experimental results on facial age estimation, aesthetic score regression, and historical color image classification show that the proposed algorithm can cluster ordered data effectively and also yield excellent rank estimation performance.

Rapid Task-Solving in Novel Environments

- Samuel Ritter, Ryan Faulkner, Laurent Sartran, Adam Santoro, Matthew Botvinick, David Raposo
- abstract@open-review(Poster): We propose the challenge of rapid task-solving in novel environments (RTS), wherein an agent must solve a series of tasks as rapidly as possible in an unfamiliar environment. An effective RTS agent must balance between exploring the unfamiliar environment and solving its current task, all while building a model of the new environment over which it can plan when faced with later tasks. While modern deep RL agents exhibit some of these abilities in isolation, none are suitable for the full RTS challenge. To enable progress toward RTS, we introduce two challenge domains: (1) a minimal RTS challenge called the Memory&Planning Game and (2) One-Shot StreetLearn Navigation, which introduces scale and complexity from real-world data. We demonstrate that state-of-the-art deep RL agents fail at RTS in both domains, and that this failure is due to an inability to plan over gathered knowledge. We develop Episodic Planning Networks (EPNs) and show that deep-RL agents with EPNs excel at RTS, outperforming the nearest baseline by factors of 2-3 and learning to navigate held-out StreetLearn maps within a single episode. We show that EPNs learn to execute a value iteration-like planning algorithm and that they generalize to situations beyond their training experience.

DINO: A Conditional Energy-Based GAN for Domain Translation

- Konstantinos Vougioukas, Stavros Petridis, Maja Pantic
- abstract@open-review(Poster): Domain translation is the process of transforming data from one domain to another while preserving the common semantics. Some of the most popular domain translation systems are based on conditional generative adversarial networks, which use source domain data to drive the generator and as an input to the discriminator. However, this approach does not enforce the preservation of shared semantics since the conditional input can often be ignored by the discriminator. We propose an alternative method for conditioning and present a new framework, where two networks are simultaneously trained, in a supervised manner, to perform domain translation in opposite directions. Our method is not only better at capturing the shared information between two domains but is more generic and can be applied to a broader range of problems. The proposed framework performs well even in challenging cross-modal translations, such as video-driven speech reconstruction, for which other systems struggle to maintain correspondence.

Removing Undesirable Feature Contributions Using Out-of-Distribution Data

- Saehyung Lee, Changhwa Park, Hyungyu Lee, Jihun Yi, Jonghyun Lee, Sungroh Yoon
- abstract@open-review(Poster): Several data augmentation methods deploy unlabeled-in-distribution (UID) data to bridge the gap between the training and inference of neural networks. However, these methods have clear limitations in terms of availability of UID data and dependence of algorithms on pseudo-labels. Herein, we propose a data augmentation method to improve generalization in both adversarial and standard learning by using out-of-distribution (OOD) data that are devoid of the abovementioned issues. We show how to improve generalization theoretically using OOD data in each learning scenario and complement our theoretical analysis with experiments on CIFAR-10, CIFAR-100, and a subset of ImageNet. The results indicate that undesirable features are shared even among image data that seem to have little correlation from a human point of view. We also present the advantages of the proposed method through comparison with other data augmentation methods, which can be used in the absence of UID data. Furthermore, we demonstrate that the proposed method can further improve the existing state-of-the-art adversarial training.

Accurate Learning of Graph Representations with Graph Multiset Pooling

- Jinheon Baek, Minki Kang, Sung Ju Hwang
- abstract@open-review(Poster): Graph neural networks have been widely used on modeling graph data, achieving impressive results on node classification and link prediction tasks. Yet, obtaining an accurate representation for a graph further requires a pooling function that maps a set of node representations into a compact form. A simple sum or average over all node representations considers all node features equally without consideration of their task relevance, and any structural dependencies among them. Recently proposed hierarchical graph pooling methods, on the other hand, may yield the same representation for two different graphs that are distinguished by the Weisfeiler-Lehman test, as they suboptimally preserve information from the node features. To tackle these limitations of existing graph pooling methods, we first formulate the graph pooling problem as a multiset encoding problem with auxiliary information about the graph structure, and propose a Graph Multiset Transformer (GMT) which is a multi-head attention based global pooling layer that captures the interaction between nodes according to their structural dependencies. We show that GMT satisfies both injectiveness and permutation invariance, such that it is at most as powerful as the Weisfeiler-Lehman graph isomorphism test. Moreover, our methods can be easily extended to the previous node clustering approaches for hierarchical graph pooling. Our experimental results show that GMT significantly outperforms state-of-the-art graph pooling methods on graph classification benchmarks with high memory and time efficiency, and obtains even larger performance gain on graph reconstruction and generation tasks.

Federated Semi-Supervised Learning with Inter-Client Consistency & Disjoint Learning

- Wonyong Jeong, Jaehong Yoon, Eunho Yang, Sung Ju Hwang
- abstract@open-review(Poster): While existing federated learning approaches mostly require that clients have fully-labeled data to train on, in realistic settings, data obtained at the client-side often comes without any accompanying labels. Such deficiency of labels may result from either high labeling cost, or difficulty of annotation due to the requirement of expert knowledge. Thus the private data at each client may be either partly labeled, or completely

unlabeled with labeled data being available only at the server, which leads us to a new practical federated learning problem, namely Federated Semi-Supervised Learning (FSSL). In this work, we study two essential scenarios of FSSL based on the location of the labeled data. The first scenario considers a conventional case where clients have both labeled and unlabeled data (labels-at-client), and the second scenario considers a more challenging case, where the labeled data is only available at the server (labels-at-server). We then propose a novel method to tackle the problems, which we refer to as Federated Matching (FedMatch). FedMatch improves upon naive combinations of federated learning and semi-supervised learning approaches with a new inter-client consistency loss and decomposition of the parameters for disjoint learning on labeled and unlabeled data. Through extensive experimental validation of our method in the two different scenarios, we show that our method outperforms both local semi-supervised learning and baselines which naively combine federated learning with semi-supervised learning.

Contrastive Learning with Adversarial Perturbations for Conditional Text Generation

- Seanie Lee, Dong Bok Lee, Sung Ju Hwang
- abstract@open-review(Poster): Recently, sequence-to-sequence (seq2seq) models with the Transformer architecture have achieved remarkable performance on various conditional text generation tasks, such as machine translation. However, most of them are trained with teacher forcing with the ground truth label given at each time step, without being exposed to incorrectly generated tokens during training, which hurts its generalization to unseen inputs, that is known as the "exposure bias" problem. In this work, we propose to solve the conditional text generation problem by contrasting positive pairs with negative pairs, such that the model is exposed to various valid or incorrect perturbations of the inputs, for improved generalization. However, training the model with naïve contrastive learning framework using random non-target sequences as negative examples is suboptimal, since they are easily distinguishable from the correct output, especially so with models pretrained with large text corpora. Also, generating positive examples requires domain-specific augmentation heuristics which may not generalize over diverse domains. To tackle this problem, we propose a principled method to generate positive and negative samples for contrastive learning of seq2seq models. Specifically, we generate negative examples by adding small perturbations to the input sequence to minimize its conditional likelihood, and positive examples by adding large perturbations while enforcing it to have a high conditional likelihood. Such `"hard" positive and negative pairs generated using our method guides the model to better distinguish correct outputs from incorrect ones. We empirically show that our proposed method significantly improves the generalization of the seq2seq on three text generation tasks --- machine translation, text summarization, and question generation.

Optimal Conversion of Conventional Artificial Neural Networks to Spiking Neural Networks

- Shikuang Deng, Shi Gu
- abstract@open-review(Poster): Spiking neural networks (SNNs) are biology-inspired artificial neural networks (ANNs) that comprise of spiking neurons to process asynchronous discrete signals. While more efficient in power consumption and inference speed on the neuromorphic hardware, SNNs are usually difficult to train directly from scratch with spikes due to the discreteness. As an alternative, many efforts have been devoted to converting conventional ANNs into SNNs by copying the weights from ANNs and adjusting the spiking threshold potential of neurons in SNNs. Researchers have designed new SNN architectures and conversion algorithms to diminish the conversion error. However, an effective conversion should address the difference between the SNN and ANN architectures with an efficient approximation of the loss function, which is missing in the field. In this work, we analyze the conversion error by recursive reduction to layer-wise summation and propose a novel strategic pipeline that transfers the weights to the target SNN by combining threshold balance and soft-reset mechanisms. This pipeline enables almost no accuracy loss between the converted SNNs and conventional ANNs with only $\sim1/10$ of the typical SNN simulation time. Our method is promising to get implanted onto embedded platforms with better support of SNNs with limited energy and memory. Codes are available at https://github.com/Jackn0/snn_optimal_conversion_pipeline.

Efficient Continual Learning with Modular Networks and Task-Driven Priors

- Tom Veniat, Ludovic Denoyer, MarcAurelio Ranzato
- abstract@open-review(Poster): Existing literature in Continual Learning (CL) has focused on overcoming catastrophic forgetting, the inability of the learner to recall how to perform tasks observed in the past. There are however other desirable properties of a CL system, such as the ability to transfer knowledge from previous tasks and to scale memory and compute sub-linearly with the number of tasks. Since most current benchmarks focus only on forgetting using short streams of tasks, we first propose a new suite of benchmarks to probe CL algorithms across these new axes. Finally, we introduce a new modular architecture, whose modules represent atomic skills that can be composed to perform a certain task. Learning a task reduces to figuring out which past modules to re-use, and which new modules to instantiate to solve the current task. Our learning algorithm leverages a task-driven prior over the exponential search space of all possible ways to combine modules, enabling efficient learning on long streams of tasks. Our experiments show that this modular architecture and learning algorithm perform competitively on widely used CL benchmarks while yielding superior performance on the more challenging benchmarks we introduce in this work. The Benchmark is publicly available at https://github.com/facebookresearch/CTrLBenchmark.

On the Universality of Rotation Equivariant Point Cloud Networks

- Nadav Dym, Haggai Maron
- abstract@open-review(Poster): Learning functions on point clouds has applications in many fields, including computer vision, computer graphics, physics, and chemistry. Recently, there has been a growing interest in neural architectures that are invariant or equivariant to all three shape-preserving transformations of point clouds: translation, rotation, and permutation. In this paper, we present a first study of the approximation power of these architectures. We first derive two sufficient conditions for an equivariant architecture to have the universal approximation property, based on a novel characterization of the space of equivariant polynomials. We then use these conditions to show that two recently suggested models, Tensor field Networks and SE3-Transformers, are universal, and for devising two other novel universal architectures.

Neural Learning of One-of-Many Solutions for Combinatorial Problems in Structured Output Spaces

- Yatin Nandwani, Deepanshu Jindal, Mausam ., Parag Singla
- abstract@open-review(Poster): Recent research has proposed neural architectures for solving combinatorial problems in structured output spaces. In many such problems, there may exist multiple solutions for a given input, e.g. a partially filled Sudoku puzzle may have many completions satisfying all constraints. Further, we are often interested in finding any "one" of the possible solutions, without any preference between them. Existing approaches completely ignore this solution multiplicity. In this paper, we argue that being oblivious to the presence of multiple solutions can severely hamper their training ability. Our contribution is two-fold. First, we formally define the task of learning one-of-many solutions for combinatorial problems in structured output spaces, which is applicable for solving several problems of interest such as N-Queens, and Sudoku. Second, we present a generic learning framework that adapts an existing prediction network for a combinatorial problem to handle solution multiplicity. Our framework uses a selection module, whose goal is to dynamically determine, for every input, the solution that is most effective for training the network parameters in any given learning iteration. We propose an RL based approach to jointly train the selection module with the prediction network. Experiments on three different domains, and using two different prediction networks, demonstrate that our framework significantly improves the accuracy in our setting, obtaining up to 21 pt gain over the baselines.

LambdaNetworks: Modeling long-range Interactions without Attention

- Irwan Bello

- abstract@open-review(Spotlight): We present lambda layers -- an alternative framework to self-attention -- for capturing long-range interactions between an input and structured contextual information (e.g. a pixel surrounded by other pixels). Lambda layers capture such interactions by transforming available contexts into linear functions, termed lambdas, and applying these linear functions to each input separately. Similar to linear attention, lambda layers bypass expensive attention maps, but in contrast, they model both content and position-based interactions which enables their application to large structured inputs such as images. The resulting neural network architectures, LambdaNetworks, significantly outperform their convolutional and attentional counterparts on ImageNet classification, COCO object detection and instance segmentation, while being more computationally efficient. Additionally, we design LambdaResNets, a family of hybrid architectures across different scales, that considerably improves the speed-accuracy tradeoff of image classification models. LambdaResNets reach excellent accuracies on ImageNet while being 3.2 - 4.4x faster than the popular EfficientNets on modern machine learning accelerators. In large-scale semi-supervised training with an additional 130M pseudo-labeled images, LambdaResNets achieve up to 86.7% ImageNet accuracy while being 9.5x faster than EfficientNet NoisyStudent and 9x faster than a Vision Transformer with comparable accuracies.

GAN2GAN: Generative Noise Learning for Blind Denoising with Single Noisy Images

- Sungmin Cha, Taeeon Park, Byeongjoon Kim, Jongduk Baek, Taesup Moon
- abstract@open-review(Poster): We tackle a challenging blind image denoising problem, in which only single distinct noisy images are available for training a denoiser, and no information about noise is known, except for it being zero-mean, additive, and independent of the clean image. In such a setting, which often occurs in practice, it is not possible to train a denoiser with the standard discriminative training or with the recently developed Noise2Noise (N2N) training; the former requires the underlying clean image for the given noisy image, and the latter requires two independently realized noisy image pair for a clean image. To that end, we propose GAN2GAN (Generated-Artificial-Noise to Generated-Artificial-Noise) method that first learns a generative model that can 1) simulate the noise in the given noisy images and 2) generate a rough, noisy estimates of the clean images, then 3) iteratively trains a denoiser with subsequently synthesized noisy image pairs (as in N2N), obtained from the generative model. In results, we show the denoiser trained with our GAN2GAN achieves an impressive denoising performance on both synthetic and real-world datasets for the blind denoising setting; it almost approaches the performance of the standard discriminatively-trained or N2N-trained models that have more information than ours, and it significantly outperforms the recent baseline for the same setting, \textit{e.g.}, Noise2Void, and a more conventional yet strong one, BM3D. The official code of our method is available at https://github.com/csm9493/GAN2GAN.

CPR: Classifier-Projection Regularization for Continual Learning

- Sungmin Cha, Hsiang Hsu, Taebaek Hwang, Flavio Calmon, Taesup Moon
- abstract@open-review(Poster): We propose a general, yet simple patch that can be applied to existing regularization-based continual learning methods called classifier-projection regularization (CPR). Inspired by both recent results on neural networks with wide local minima and information theory, CPR adds an additional regularization term that maximizes the entropy of a classifier's output probability. We demonstrate that this additional term can be interpreted as a projection of the conditional probability given by a classifier's output to the uniform distribution. By applying the Pythagorean theorem for KL divergence, we then prove that this projection may (in theory) improve the performance of continual learning methods. In our extensive experimental results, we apply CPR to several state-of-the-art regularization-based continual learning methods and benchmark performance on popular image recognition datasets. Our results demonstrate that CPR indeed promotes a wide local minima and significantly improves both accuracy and plasticity while simultaneously mitigating the catastrophic forgetting of baseline continual learning methods. The codes and scripts for this work are available at https://github.com/csm9493/CPR_CL.

Contrastive Divergence Learning is a Time Reversal Adversarial Game

- Omer Yair, Tomer Michaeli
- abstract@open-review(Spotlight): Contrastive divergence (CD) learning is a classical method for fitting unnormalized statistical models to data samples. Despite its wide-spread use, the convergence properties of this algorithm are still not well understood. The main source of difficulty is an unjustified approximation which has been used to derive the gradient of the loss. In this paper, we present an alternative derivation of CD that does not require any approximation and sheds new light on the objective that is actually being optimized by the algorithm. Specifically, we show that CD is an adversarial learning procedure, where a discriminator attempts to classify whether a Markov chain generated from the model has been time-reversed. Thus, although predating generative adversarial networks (GANs) by more than a decade, CD is, in fact, closely related to these techniques. Our derivation settles well with previous observations, which have concluded that CD's update steps cannot be expressed as the gradients of any fixed objective function. In addition, as a byproduct, our derivation reveals a simple correction that can be used as an alternative to Metropolis-Hastings rejection, which is required when the underlying Markov chain is inexact (e.g., when using Langevin dynamics with a large step).

On the Dynamics of Training Attention Models

- Haoye Lu, Yongyi Mao, Amiya Nayak
- abstract@open-review(Poster): The attention mechanism has been widely used in deep neural networks as a model component. By now, it has become a critical building block in many state-of-the-art natural language models. Despite its great success established empirically, the working mechanism of attention has not been investigated at a sufficient theoretical depth to date. In this paper, we set up a simple text classification task and study the dynamics of training a simple attention-based classification model using gradient descent. In this setting, we show that, for the discriminative words that the model should attend to, a persisting identity exists relating its embedding and the inner product of its key and the query. This allows us to prove that training must converge to attending to the discriminative words when the attention output is classified by a linear classifier. Experiments are performed, which validate our theoretical analysis and provide further insights.

Model-Based Offline Planning

- Arthur Argenson, Gabriel Dulac-Arnold
- abstract@open-review(Poster): Offline learning is a key part of making reinforcement learning (RL) useable in real systems. Offline RL looks at scenarios where there is data from a system's operation, but no direct access to the system when learning a policy. Recent work on training RL policies from offline data has shown results both with model-free policies learned directly from the data, or with planning on top of learnt models of the data. Model-free policies tend to be more performant, but are more opaque, harder to command externally, and less easy to integrate into larger systems. We propose an offline learner that generates a model that can be used to control the system directly through planning. This allows us to have easily controllable policies directly from data, without ever interacting with the system. We show the performance of our algorithm, Model-Based Offline Planning (MBOP) on a series of robotics-inspired tasks, and demonstrate its ability leverage planning to respect environmental constraints. We are able to find near-optimal polices for certain simulated systems from as little as 50 seconds of real-time system interaction, and create zero-shot goal-conditioned policies on a series of environments.

Modelling Hierarchical Structure between Dialogue Policy and Natural Language Generator with Option Framework for Task-oriented Dialogue System

- Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, Yunjie Gu

- abstract@open-review(Poster): Designing task-oriented dialogue systems is a challenging research topic, since it needs not only to generate utterances fulfilling user requests but also to guarantee the comprehensibility. Many previous works trained end-to-end (E2E) models with supervised learning (SL), however, the bias in annotated system utterances remains as a bottleneck. Reinforcement learning (RL) deals with the problem through using non-differentiable evaluation metrics (e.g., the success rate) as rewards. Nonetheless, existing works with RL showed that the comprehensibility of generated system utterances could be corrupted when improving the performance on fulfilling user requests. In our work, we (1) propose modelling the hierarchical structure between dialogue policy and natural language generator (NLG) with the option framework, called HDNO, where the latent dialogue act is applied to avoid designing specific dialogue act representations; (2) train HDNO via hierarchical reinforcement learning (HRL), as well as suggest the asynchronous updates between dialogue policy and NLG during training to theoretically guarantee their convergence to a local maximizer; and (3) propose using a discriminator modelled with language models as an additional reward to further improve the comprehensibility. We test HDNO on MultiWoz 2.0 and MultiWoz 2.1, the datasets on multi-domain dialogues, in comparison with word-level E2E model trained with RL, LaRL and HDSA, showing improvements on the performance evaluated by automatic evaluation metrics and human evaluation. Finally, we demonstrate the semantic meanings of latent dialogue acts to show the explanability for HDNO.

Neural Synthesis of Binaural Speech From Mono Audio

- Alexander Richard, Dejan Markovic, Israel D. Gebru, Steven Krenn, Gladstone Alexander Butler, Fernando Torre, Yaser Sheikh
- abstract@open-review(Oral): We present a neural rendering approach for binaural sound synthesis that can produce realistic and spatially accurate binaural sound in realtime. The network takes, as input, a single-channel audio source and synthesizes, as output, two-channel binaural sound, conditioned on the relative position and orientation of the listener with respect to the source. We investigate deficiencies of the l2-loss on raw waveforms in a theoretical analysis and introduce an improved loss that overcomes these limitations. In an empirical evaluation, we establish that our approach is the first to generate spatially accurate waveform outputs (as measured by real recordings) and outperforms existing approaches by a considerable margin, both quantitatively and in a perceptual study. Dataset and code are available online.

Distilling Knowledge from Reader to Retriever for Question Answering

- Gautier Izacard, Edouard Grave
- abstract@open-review(Poster): The task of information retrieval is an important component of many natural language processing systems, such as open domain question answering. While traditional methods were based on hand-crafted features, continuous representations based on neural networks recently obtained competitive results. A challenge of using such methods is to obtain supervised data to train the retriever model, corresponding to pairs of query and support documents. In this paper, we propose a technique to learn retriever models for downstream tasks, inspired by knowledge distillation, and which does not require annotated pairs of query and documents. Our approach leverages attention scores of a reader model, used to solve the task based on retrieved documents, to obtain synthetic labels for the retriever. We evaluate our method on question answering, obtaining state-of-the-art results.

Efficient Certified Defenses Against Patch Attacks on Image Classifiers

- Jan Hendrik Metzen, Maksym Yatsura
- abstract@open-review(Poster): Adversarial patches pose a realistic threat model for physical world attacks on autonomous systems via their perception component. Autonomous systems in safety-critical domains such as automated driving should thus contain a fail-safe fallback component that combines certifiable robustness against patches with efficient inference while maintaining high performance on clean inputs. We propose BagCert, a novel combination of model architecture and certification procedure that allows efficient certification. We derive a loss that enables end-to-end optimization of certified robustness against patches of different sizes and locations. On CIFAR10, BagCert certifies 10.000 examples in 43 seconds on a single GPU and obtains 86% clean and 60% certified accuracy against 5x5 patches.

Quantifying Differences in Reward Functions

- Adam Gleave, Michael D Dennis, Shane Legg, Stuart Russell, Jan Leike
- abstract@open-review(Spotlight): For many tasks, the reward function is inaccessible to introspection or too complex to be specified procedurally, and must instead be learned from user data. Prior work has evaluated learned reward functions by evaluating policies optimized for the learned reward. However, this method cannot distinguish between the learned reward function failing to reflect user preferences and the policy optimization process failing to optimize the learned reward. Moreover, this method can only tell us about behavior in the evaluation environment, but the reward may incentivize very different behavior in even a slightly different deployment environment. To address these problems, we introduce the Equivalent-Policy Invariant Comparison (EPIC) distance to quantify the difference between two reward functions directly, without a policy optimization step. We prove EPIC is invariant on an equivalence class of reward functions that always induce the same optimal policy. Furthermore, we find EPIC can be efficiently approximated and is more robust than baselines to the choice of coverage distribution. Finally, we show that EPIC distance bounds the regret of optimal policies even under different transition dynamics, and we confirm empirically that it predicts policy training success. Our source code is available at https://github.com/HumanCompatibleAI/evaluating-rewards.

Dataset Condensation with Gradient Matching

- Bo Zhao, Konda Reddy Mopuri, Hakan Bilen
- abstract@open-review(Oral): As the state-of-the-art machine learning methods in many fields rely on larger datasets, storing datasets and training models on them become significantly more expensive. This paper proposes a training set synthesis technique for data-efficient learning, called Dataset Condensation, that learns to condense large dataset into a small set of informative synthetic samples for training deep neural networks from scratch. We formulate this goal as a gradient matching problem between the gradients of deep neural network weights that are trained on the original and our synthetic data. We rigorously evaluate its performance in several computer vision benchmarks and demonstrate that it significantly outperforms the state-of-the-art methods. Finally we explore the use of our method in continual learning and neural architecture search and report promising gains when limited memory and computations are available.

Taking Notes on the Fly Helps Language Pre-Training

- Qiyu Wu, Chen Xing, Yatao Li, Guolin Ke, Di He, Tie-Yan Liu
- abstract@open-review(Poster): How to make unsupervised language pre-training more efficient and less resource-intensive is an important research direction in NLP. In this paper, we focus on improving the efficiency of language pre-training methods through providing better data utilization. It is well-known that in language data corpus, words follow a heavy-tail distribution. A large proportion of words appear only very few times and the embeddings of rare words are usually poorly optimized. We argue that such embeddings carry inadequate semantic signals, which could make the data utilization inefficient and slow down the pre-training of the entire model. To mitigate this problem, we propose Taking Notes on the Fly (TNF), which takes notes for rare words on the fly during pre-training to help the model understand them when they occur next time. Specifically, TNF maintains a note dictionary and saves a rare word's contextual information in it as notes when the rare word occurs in a sentence. When the same rare word occurs again during training, the note information saved beforehand can be employed to enhance the semantics of the current sentence. By doing so, TNF provides a better data utilization since cross-sentence information is employed to cover the inadequate semantics caused by rare words in the sentences. We implement TNF on both BERT and ELECTRA to check its efficiency and effectiveness. Experimental results show that TNF's training time is 60% less than its backbone pre-

training models when reaching the same performance. When trained with same number of iterations, TNF outperforms its backbone methods on most of downstream tasks and the average GLUE score. Code is attached in the supplementary material.

Graph Edit Networks

- Benjamin Paassen, Daniele Grattarola, Daniele Zambon, Cesare Alippi, Barbara Eva Hammer
- abstract@open-review(Poster): While graph neural networks have made impressive progress in classification and regression, few approaches to date perform time series prediction on graphs, and those that do are mostly limited to edge changes. We suggest that graph edits are a more natural interface for graph-to-graph learning. In particular, graph edits are general enough to describe any graph-to-graph change, not only edge changes; they are sparse, making them easier to understand for humans and more efficient computationally; and they are local, avoiding the need for pooling layers in graph neural networks. In this paper, we propose a novel output layer - the graph edit network - which takes node embeddings as input and generates a sequence of graph edits that transform the input graph to the output graph. We prove that a mapping between the node sets of two graphs is sufficient to construct training data for a graph edit network and that an optimal mapping yields edit scripts that are almost as short as the graph edit distance between the graphs. We further provide a proof-of-concept empirical evaluation on several graph dynamical systems, which are difficult to learn for baselines from the literature.

FOCAL: Efficient Fully-Offline Meta-Reinforcement Learning via Distance Metric Learning and Behavior Regularization

- Lanqing Li, Rui Yang, Dijun Luo
- abstract@open-review(Poster): We study the offline meta-reinforcement learning (OMRL) problem, a paradigm which enables reinforcement learning (RL) algorithms to quickly adapt to unseen tasks without any interactions with the environments, making RL truly practical in many real-world applications. This problem is still not fully understood, for which two major challenges need to be addressed. First, offline RL usually suffers from bootstrapping errors of out-of-distribution state-actions which leads to divergence of value functions. Second, meta-RL requires efficient and robust task inference learned jointly with control policy. In this work, we enforce behavior regularization on learned policy as a general approach to offline RL, combined with a deterministic context encoder for efficient task inference. We propose a novel negative-power distance metric on bounded context embedding space, whose gradients propagation is detached from the Bellman backup. We provide analysis and insight showing that some simple design choices can yield substantial improvements over recent approaches involving meta-RL and distance metric learning. To the best of our knowledge, our method is the first model-free and end-to-end OMRL algorithm, which is computationally efficient and demonstrated to outperform prior algorithms on several meta-RL benchmarks.

Effective and Efficient Vote Attack on Capsule Networks

- Jindong Gu, Baoyuan Wu, Volker Tresp
- abstract@open-review(Poster): Standard Convolutional Neural Networks (CNNs) can be easily fooled by images with small quasi-imperceptible artificial perturbations. As alternatives to CNNs, the recently proposed Capsule Networks (CapsNets) are shown to be more robust to white-box attack than CNNs under popular attack protocols. Besides, the class-conditional reconstruction part of CapsNets is also used to detect adversarial examples. In this work, we investigate the adversarial robustness of CapsNets, especially how the inner workings of CapsNets change when the output capsules are attacked. The first observation is that adversarial examples misled CapsNets by manipulating the votes from primary capsules. Another observation is the high computational cost, when we directly apply multi-step attack methods designed for CNNs to attack CapsNets, due to the computationally expensive routing mechanism. Motivated by these two observations, we propose a novel vote attack where we attack votes of CapsNets directly. Our vote attack is not only effective, but also efficient by circumventing the routing process. Furthermore, we integrate our vote attack into the detection-aware attack paradigm, which can successfully bypass the class-conditional reconstruction based detection method. Extensive experiments demonstrate the superior attack performance of our vote attack on CapsNets.

Rapid Neural Architecture Search by Learning to Generate Graphs from Datasets

- Hayeon Lee, Eunyoung Hyung, Sung Ju Hwang
- abstract@open-review(Poster): Despite the success of recent Neural Architecture Search (NAS) methods on various tasks which have shown to output networks that largely outperform human-designed networks, conventional NAS methods have mostly tackled the optimization of searching for the network architecture for a single task (dataset), which does not generalize well across multiple tasks (datasets). Moreover, since such task-specific methods search for a neural architecture from scratch for every given task, they incur a large computational cost, which is problematic when the time and monetary budget are limited. In this paper, we propose an efficient NAS framework that is trained once on a database consisting of datasets and pretrained networks and can rapidly search for a neural architecture for a novel dataset. The proposed MetaD2A (Meta Dataset-to-Architecture) model can stochastically generate graphs (architectures) from a given set (dataset) via a cross-modal latent space learned with amortized meta-learning. Moreover, we also propose a meta-performance predictor to estimate and select the best architecture without direct training on target datasets. The experimental results demonstrate that our model meta-learned on subsets of ImageNet-1K and architectures from NAS-Bench 201 search space successfully generalizes to multiple unseen datasets including CIFAR-10 and CIFAR-100, with an average search time of 33 GPU seconds. Even under MobileNetV3 search space, MetaD2A is 5.5K times faster than NSGANetV2, a transferable NAS method, with comparable performance. We believe that the MetaD2A proposes a new research direction for rapid NAS as well as ways to utilize the knowledge from rich databases of datasets and architectures accumulated over the past years. Code is available at https://github.com/HayeonLee/MetaD2A.

Impact of Representation Learning in Linear Bandits

- Jiaqi Yang, Wei Hu, Jason D. Lee, Simon Shaolei Du
- abstract@open-review(Poster): We study how representation learning can improve the efficiency of bandit problems. We study the setting where we play T linear bandits with dimension d concurrently, and these T bandit tasks share a common $k (\ll d)$ dimensional linear representation. For the finite-action setting, we present a new algorithm which achieves $\widetilde{O}(T\sqrt{kN} + \sqrt{dkNT})$ regret, where N is the number of rounds we play for each bandit. When T is sufficiently large, our algorithm significantly outperforms the naive algorithm (playing T bandits independently) that achieves $\widetilde{O}(T\sqrt{d N})$ regret. We also provide an $\Omega(T\sqrt{kN} + \sqrt{dkNT})$ regret lower bound, showing that our algorithm is minimax-optimal up to poly-logarithmic factors. Furthermore, we extend our algorithm to the infinite-action setting and obtain a corresponding regret bound which demonstrates the benefit of representation learning in certain regimes. We also present experiments on synthetic and real-world data to illustrate our theoretical findings and demonstrate the effectiveness of our proposed algorithms.

Creative Sketch Generation

- Songwei Ge, Vedanuj Goswami, Larry Zitnick, Devi Parikh
- abstract@open-review(Poster): Sketching or doodling is a popular creative activity that people engage in. However, most existing work in automatic sketch understanding or generation has focused on sketches that are quite mundane. In this work, we introduce two datasets of creative sketches -- Creative Birds and Creative Creatures -- containing 10k sketches each along with part annotations. We propose DoodlerGAN -- a part-based Generative Adversarial Network (GAN) -- to generate unseen compositions of novel part appearances. Quantitative evaluations as well as human studies demonstrate that sketches generated by our approach are more creative and of higher quality than existing approaches. In fact, in Creative Birds, subjects prefer sketches generated by DoodlerGAN over those drawn by humans!

Self-supervised Representation Learning with Relative Predictive Coding

- Yao-Hung Hubert Tsai, Martin Q. Ma, Muqiao Yang, Han Zhao, Louis-Philippe Morency, Ruslan Salakhutdinov
- abstract@open-review(Poster): This paper introduces Relative Predictive Coding (RPC), a new contrastive representation learning objective that maintains a good balance among training stability, minibatch size sensitivity, and downstream task performance. The key to the success of RPC is two-fold. First, RPC introduces the relative parameters to regularize the objective for boundedness and low variance. Second, RPC contains no logarithm and exponential score functions, which are the main cause of training instability in prior contrastive objectives. We empirically verify the effectiveness of RPC on benchmark vision and speech self-supervised learning tasks. Lastly, we relate RPC with mutual information (MI) estimation, showing RPC can be used to estimate MI with low variance.

One Network Fits All? Modular versus Monolithic Task Formulations in Neural Networks

- Atish Agarwala, Abhimanyu Das, Brendan Juba, Rina Panigrahy, Vatsal Sharan, Xin Wang, Qiuyi Zhang
- abstract@open-review(Poster): Can deep learning solve multiple, very different tasks simultaneously? We investigate how the representations of the underlying tasks affect the ability of a single neural network to learn them jointly. We present theoretical and empirical findings that a single neural network is capable of simultaneously learning multiple tasks from a combined data set, for a variety of methods for representing tasks---for example, when the distinct tasks are encoded by well-separated clusters or decision trees over some task-code attributes. Indeed, more strongly, we present a novel analysis that shows that families of simple programming-like constructs for the codes encoding the tasks are learnable by two-layer neural networks with standard training. We study more generally how the complexity of learning such combined tasks grows with the complexity of the task codes; we find that learning many tasks can be provably hard, even though the individual tasks are easy to learn. We provide empirical support for the usefulness of the learning bounds by training networks on clusters, decision trees, and SQL-style aggregation.

Conditionally Adaptive Multi-Task Learning: Improving Transfer Learning in NLP Using Fewer Parameters & Less Data

- Jonathan Pilault, Amine El hattami, Christopher Pal
- abstract@open-review(Poster): Multi-Task Learning (MTL) networks have emerged as a promising method for transferring learned knowledge across different tasks. However, MTL must deal with challenges such as: overfitting to low resource tasks, catastrophic forgetting, and negative task transfer, or learning interference. Often, in Natural Language Processing (NLP), a separate model per task is needed to obtain the best performance. However, many fine-tuning approaches are both parameter inefficient, i.e., potentially involving one new model per task, and highly susceptible to losing knowledge acquired during pretraining. We propose a novel Transformer based Hypernetwork Adapter consisting of a new conditional attention mechanism as well as a set of task-conditioned modules that facilitate weight sharing. Through this construction, we achieve more efficient parameter sharing and mitigate forgetting by keeping half of the weights of a pretrained model fixed. We also use a new multi-task data sampling strategy to mitigate the negative effects of data imbalance across tasks. Using this approach, we are able to surpass single task fine-tuning methods while being parameter and data efficient (using around 66% of the data). Compared to other BERT Large methods on GLUE, our 8-task model surpasses other Adapter methods by 2.8% and our 24-task model outperforms by 0.7-1.0% models that use MTL and single task fine-tuning. We show that a larger variant of our single multi-task model approach performs competitively across 26 NLP tasks and yields state-of-the-art results on a number of test and development sets.

A Universal Representation Transformer Layer for Few-Shot Image Classification

- Lu Liu, William L. Hamilton, Guodong Long, Jing Jiang, Hugo Larochelle
- abstract@open-review(Poster): Few-shot classification aims to recognize unseen classes when presented with only a small number of samples. We consider the problem of multi-domain few-shot image classification, where unseen classes and examples come from diverse data sources. This problem has seen growing interest and has inspired the development of benchmarks such as Meta-Dataset. A key challenge in this multi-domain setting is to effectively integrate the feature representations from the diverse set of training domains. Here, we propose a Universal Representation Transformer (URT) layer, that meta-learns to leverage universal features for few-shot classification by dynamically re-weighting and composing the most appropriate domain-specific representations. In experiments, we show that URT sets a new state-of-the-art result on Meta-Dataset. Specifically, it achieves top-performance on the highest number of data sources compared to competing methods. We analyze variants of URT and present a visualization of the attention score heatmaps that sheds light on how the model performs cross-domain generalization.

Isometric Propagation Network for Generalized Zero-shot Learning

- Lu Liu, Tianyi Zhou, Guodong Long, Jing Jiang, Xuanyi Dong, Chengqi Zhang
- abstract@open-review(Poster): Zero-shot learning (ZSL) aims to classify images of an unseen class only based on a few attributes describing that class but no access to any training sample. A popular strategy is to learn a mapping between the semantic space of class attributes and the visual space of images based on the seen classes and their data. Thus, an unseen class image can be ideally mapped to its corresponding class attributes. The key challenge is how to align the representations in the two spaces. For most ZSL settings, the attributes for each seen/unseen class are only represented by a vector while the seen-class data provide much more information. Thus, the imbalanced supervision from the semantic and the visual space can make the learned mapping easily overfitting to the seen classes. To resolve this problem, we propose Isometric Propagation Network (IPN), which learns to strengthen the relation between classes within each space and align the class dependency in the two spaces. Specifically, IPN learns to propagate the class representations on an auto-generated graph within each space. In contrast to only aligning the resulted static representation, we regularize the two dynamic propagation procedures to be isometric in terms of the two graphs' edge weights per step by minimizing a consistency loss between them. IPN achieves state-of-the-art performance on three popular ZSL benchmarks. To evaluate the generalization capability of IPN, we further build two larger benchmarks with more diverse unseen classes and demonstrate the advantages of IPN on them.

Scalable Learning and MAP Inference for Nonsymmetric Determinantal Point Processes

- Mike Gartrell, Insu Han, Elvis Dohmatob, Jennifer Gillenwater, Victor-Emmanuel Brunel
- abstract@open-review(Oral): Determinantal point processes (DPPs) have attracted significant attention in machine learning for their ability to model subsets drawn from a large item collection. Recent work shows that nonsymmetric DPP (NDPP) kernels have significant advantages over symmetric kernels in terms of modeling power and predictive performance. However, for an item collection of size M, existing NDPP learning and inference algorithms require memory quadratic in M and runtime cubic (for learning) or quadratic (for inference) in M, making them impractical for many typical subset selection tasks. In this work, we develop a learning algorithm with space and time requirements linear in M by introducing a new NDPP kernel decomposition. We also derive a linear-complexity NDPP maximum a posteriori (MAP) inference algorithm that applies not only to our new kernel but also to that of prior work. Through evaluation on real-world datasets, we show that our algorithms scale significantly better, and can match the predictive performance of prior work.

Long-tail learning via logit adjustment

- Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, Sanjiv Kumar
- abstract@open-review(Spotlight): Real-world classification problems typically exhibit an imbalanced or long-tailed label distribution, wherein many labels have only a few associated samples. This poses a challenge for generalisation on such labels, and also makes naive learning biased towards

dominant labels. In this paper, we present a statistical framework that unifies and generalises several recent proposals to cope with these challenges. Our framework revisits the classic idea of logit adjustment based on the label frequencies, which encourages a large relative margin between logits of rare positive versus dominant negative labels. This yields two techniques for long-tail learning, where such adjustment is either applied post-hoc to a trained model, or enforced in the loss during training. These techniques are statistically grounded, and practically effective on four real-world datasets with long-tailed label distributions.

Locally Free Weight Sharing for Network Width Search

- Xiu Su, Shan You, Tao Huang, Fei Wang, Chen Qian, Changshui Zhang, Chang Xu
- abstract@open-review(Spotlight): Searching for network width is an effective way to slim deep neural networks with hardware budgets. With this aim, a one-shot supernet is usually leveraged as a performance evaluator to rank the performance \wrt~different width. Nevertheless, current methods mainly follow a manually fixed weight sharing pattern, which is limited to distinguish the performance gap of different width. In this paper, to better evaluate each width, we propose a locally free weight sharing strategy (CafeNet) accordingly. In CafeNet, weights are more freely shared, and each width is jointly indicated by its base channels and free channels, where free channels are supposed to locate freely in a local zone to better represent each width. Besides, we propose to further reduce the search space by leveraging our introduced FLOPs-sensitive bins. As a result, our CafeNet can be trained stochastically and get optimized within a min-min strategy. Extensive experiments on ImageNet, CIFAR-10, CelebA and MS COCO dataset have verified our superiority comparing to other state-of-the-art baselines. For example, our method can further boost the benchmark NAS network EfficientNet-B0 by 0.41\% via searching its width more delicately.

Mutual Information State Intrinsic Control

- Rui Zhao, Yang Gao, Pieter Abbeel, Volker Tresp, Wei Xu
- abstract@open-review(Spotlight): Reinforcement learning has been shown to be highly successful at many challenging tasks. However, success heavily relies on well-shaped rewards. Intrinsically motivated RL attempts to remove this constraint by defining an intrinsic reward function. Motivated by the self-consciousness concept in psychology, we make a natural assumption that the agent knows what constitutes itself, and propose a new intrinsic objective that encourages the agent to have maximum control on the environment. We mathematically formalize this reward as the mutual information between the agent state and the surrounding state under the current agent policy. With this new intrinsic motivation, we are able to outperform previous methods, including being able to complete the pick-and-place task for the first time without using any task reward. A video showing experimental results is available at https://youtu.be/AUCwc9RThpk.

Multivariate Probabilistic Time Series Forecasting via Conditioned Normalizing Flows

- Kashif Rasul, Abdul-Saboor Sheikh, Ingmar Schuster, Urs M Bergmann, Roland Vollgraf
- abstract@open-review(Spotlight): Time series forecasting is often fundamental to scientific and engineering problems and enables decision making. With ever increasing data set sizes, a trivial solution to scale up predictions is to assume independence between interacting time series. However, modeling statistical dependencies can improve accuracy and enable analysis of interaction effects. Deep learning methods are well suited for this problem, but multi-variate models often assume a simple parametric distribution and do not scale to high dimensions. In this work we model the multi-variate temporal dynamics of time series via an autoregressive deep learning model, where the data distribution is represented by a conditioned normalizing flow. This combination retains the power of autoregressive models, such as good performance in extrapolation into the future, with the flexibility of flows as a general purpose high-dimensional distribution model, while remaining computationally tractable. We show that it improves over the state-of-the-art for standard metrics on many real-world data sets with several thousand interacting time-series.

Geometry-aware Instance-reweighted Adversarial Training

- Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, Mohan Kankanhalli
- abstract@open-review(Oral): In adversarial machine learning, there was a common belief that robustness and accuracy hurt each other. The belief was challenged by recent studies where we can maintain the robustness and improve the accuracy. However, the other direction, whether we can keep the accuracy and improve the robustness, is conceptually and practically more interesting, since robust accuracy should be lower than standard accuracy for any model. In this paper, we show this direction is also promising. Firstly, we find even over-parameterized deep networks may still have insufficient model capacity, because adversarial training has an overwhelming smoothing effect. Secondly, given limited model capacity, we argue adversarial data should have unequal importance: geometrically speaking, a natural data point closer to/farther from the class boundary is less/more robust, and the corresponding adversarial data point should be assigned with larger/smaller weight. Finally, to implement the idea, we propose geometry-aware instance-reweighted adversarial training, where the weights are based on how difficult it is to attack a natural data point. Experiments show that our proposal boosts the robustness of standard adversarial training; combining two directions, we improve both robustness and accuracy of standard adversarial training.

Towards Impartial Multi-task Learning

- Liyang Liu, Yi Li, Zhanghui Kuang, Jing-Hao Xue, Yimin Chen, Wenming Yang, Qingmin Liao, Wayne Zhang
- abstract@open-review(Poster): Multi-task learning (MTL) has been widely used in representation learning. However, naively training all tasks simultaneously may lead to the partial training issue, where specific tasks are trained more adequately than others. In this paper, we propose to learn multiple tasks impartially. Specifically, for the task-shared parameters, we optimize the scaling factors via a closed-form solution, such that the aggregated gradient (sum of raw gradients weighted by the scaling factors) has equal projections onto individual tasks. For the task-specific parameters, we dynamically weigh the task losses so that all of them are kept at a comparable scale. Further, we find the above gradient balance and loss balance are complementary and thus propose a hybrid balance method to further improve the performance. Our impartial multi-task learning (IMTL) can be end-to-end trained without any heuristic hyper-parameter tuning, and is general to be applied on all kinds of losses without any distribution assumption. Moreover, our IMTL can converge to similar results even when the task losses are designed to have different scales, and thus it is scale-invariant. We extensively evaluate our IMTL on the standard MTL benchmarks including Cityscapes, NYUv2 and CelebA. It outperforms existing loss weighting methods under the same experimental settings.

On Learning Universal Representations Across Languages

- Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, Weihua Luo
- abstract@open-review(Poster): Recent studies have demonstrated the overwhelming advantage of cross-lingual pre-trained models (PTMs), such as multilingual BERT and XLM, on cross-lingual NLP tasks. However, existing approaches essentially capture the co-occurrence among tokens through involving the masked language model (MLM) objective with token-level cross entropy. In this work, we extend these approaches to learn sentence-level representations and show the effectiveness on cross-lingual understanding and generation. Specifically, we propose a Hierarchical Contrastive Learning (HiCTL) method to (1) learn universal representations for parallel sentences distributed in one or multiple languages and (2) distinguish the semantically-related words from a shared cross-lingual vocabulary for each sentence. We conduct evaluations on two challenging cross-lingual tasks, XTREME and machine translation. Experimental results show that the HiCTL outperforms the state-of-the-art XLM-R by an absolute gain of 4.2% accuracy on the XTREME benchmark as well as achieves substantial improvements on both of the high resource and low-resource English$\rightarrow$X translation tasks over strong baselines.

Isotropy in the Contextual Embedding Space: Clusters and Manifolds

- Xingyu Cai, Jiaji Huang, Yuchen Bian, Kenneth Church
- abstract@open-review(Poster): The geometric properties of contextual embedding spaces for deep language models such as BERT and ERNIE, have attracted considerable attention in recent years. Investigations on the contextual embeddings demonstrate a strong anisotropic space such that most of the vectors fall within a narrow cone, leading to high cosine similarities. It is surprising that these LMs are as successful as they are, given that most of their embedding vectors are as similar to one another as they are. In this paper, we argue that the isotropy indeed exists in the space, from a different but more constructive perspective. We identify isolated clusters and low dimensional manifolds in the contextual embedding space, and introduce tools to both qualitatively and quantitatively analyze them. We hope the study in this paper could provide insights towards a better understanding of the deep language models.

MoPro: Webly Supervised Learning with Momentum Prototypes

- Junnan Li, Caiming Xiong, Steven Hoi
- abstract@open-review(Poster): We propose a webly-supervised representation learning method that does not suffer from the annotation unscalability of supervised learning, nor the computation unscalability of self-supervised learning. Most existing works on webly-supervised representation learning adopt a vanilla supervised learning method without accounting for the prevalent noise in the training data, whereas most prior methods in learning with label noise are less effective for real-world large-scale noisy data. We propose momentum prototypes (MoPro), a simple contrastive learning method that achieves online label noise correction, out-of-distribution sample removal, and representation learning. MoPro achieves state-of-the-art performance on WebVision, a weakly-labeled noisy dataset. MoPro also shows superior performance when the pretrained model is transferred to down-stream image classification and detection tasks. It outperforms the ImageNet supervised pretrained model by +10.5 on 1-shot classification on VOC, and outperforms the best self-supervised pretrained model by +17.3 when finetuned on 1% of ImageNet labeled samples. Furthermore, MoPro is more robust to distribution shifts. Code and pretrained models are available at https://github.com/salesforce/MoPro.

GraphCodeBERT: Pre-training Code Representations with Data Flow

- Daya Guo, Shuo Ren, Shuai Lu, Zhangyin Feng, Duyu Tang, Shujie LIU, Long Zhou, Nan Duan, Alexey Svyatkovskiy, Shengyu Fu, Michele Tufano, Shao Kun Deng, Colin Clement, Dawn Drain, Neel Sundaresan, Jian Yin, Daxin Jiang, Ming Zhou
- abstract@open-review(Poster): Pre-trained models for programming language have achieved dramatic empirical improvements on a variety of code-related tasks such as code search, code completion, code summarization, etc. However, existing pre-trained models regard a code snippet as a sequence of tokens, while ignoring the inherent structure of code, which provides crucial code semantics and would enhance the code understanding process. We present GraphCodeBERT, a pre-trained model for programming language that considers the inherent structure of code. Instead of taking syntactic-level structure of code like abstract syntax tree (AST), we use data flow in the pre-training stage, which is a semantic-level structure of code that encodes the relation of "where-the-value-comes-from" between variables. Such a semantic-level structure is neat and does not bring an unnecessarily deep hierarchy of AST, the property of which makes the model more efficient. We develop GraphCodeBERT based on Transformer. In addition to using the task of masked language modeling, we introduce two structure-aware pre-training tasks. One is to predict code structure edges, and the other is to align representations between source code and code structure. We implement the model in an efficient way with a graph-guided masked attention function to incorporate the code structure. We evaluate our model on four tasks, including code search, clone detection, code translation, and code refinement. Results show that code structure and newly introduced pre-training tasks can improve GraphCodeBERT and achieves state-of-the-art performance on the four downstream tasks. We further show that the model prefers structure-level attentions over token-level attentions in the task of code search.

Enjoy Your Editing: Controllable GANs for Image Editing via Latent Space Navigation

- Peiye Zhuang, Oluwasanmi O Koyejo, Alex Schwing
- abstract@open-review(Poster): Controllable semantic image editing enables a user to change entire image attributes with a few clicks, e.g., gradually making a summer scene look like it was taken in winter. Classic approaches for this task use a Generative Adversarial Net (GAN) to learn a latent space and suitable latent-space transformations. However, current approaches often suffer from attribute edits that are entangled, global image identity changes, and diminished photo-realism. To address these concerns, we learn multiple attribute transformations simultaneously, integrate attribute regression into the training of transformation functions, and apply a content loss and an adversarial loss that encourages the maintenance of image identity and photo-realism. We propose quantitative evaluation strategies for measuring controllable editing performance, unlike prior work, which primarily focuses on qualitative evaluation. Our model permits better control for both single- and multiple-attribute editing while preserving image identity and realism during transformation. We provide empirical results for both natural and synthetic images, highlighting that our model achieves state-of-the-art performance for targeted image manipulation.

Fourier Neural Operator for Parametric Partial Differential Equations

- Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede liu, Kaushik Bhattacharya, Andrew Stuart, Anima Anandkumar
- abstract@open-review(Poster): The classical development of neural networks has primarily focused on learning mappings between finite-dimensional Euclidean spaces. Recently, this has been generalized to neural operators that learn mappings between function spaces. For partial differential equations (PDEs), neural operators directly learn the mapping from any functional parametric dependence to the solution. Thus, they learn an entire family of PDEs, in contrast to classical methods which solve one instance of the equation. In this work, we formulate a new neural operator by parameterizing the integral kernel directly in Fourier space, allowing for an expressive and efficient architecture. We perform experiments on Burgers' equation, Darcy flow, and Navier-Stokes equation. The Fourier neural operator is the first ML-based method to successfully model turbulent flows with zero-shot super-resolution. It is up to three orders of magnitude faster compared to traditional PDE solvers. Additionally, it achieves superior accuracy compared to previous learning-based solvers under fixed resolution.

Training BatchNorm and Only BatchNorm: On the Expressive Power of Random Features in CNNs

- Jonathan Frankle, David J. Schwab, Ari S. Morcos
- abstract@open-review(Poster): A wide variety of deep learning techniques from style transfer to multitask learning rely on training affine transformations of features. Most prominent among these is the popular feature normalization technique BatchNorm, which normalizes activations and then subsequently applies a learned affine transform. In this paper, we aim to understand the role and expressive power of affine parameters used to transform features in this way. To isolate the contribution of these parameters from that of the learned features they transform, we investigate the performance achieved when training only these parameters in BatchNorm and freezing all weights at their random initializations. Doing so leads to surprisingly high performance considering the significant limitations that this style of training imposes. For example, sufficiently deep ResNets reach 82% (CIFAR-10) and 32% (ImageNet, top-5) accuracy in this configuration, far higher than when training an equivalent number of randomly chosen parameters elsewhere in the network. BatchNorm achieves this performance in part by naturally learning to disable around a third of the random features. Not only do these results highlight the expressive power of affine parameters in deep learning, but - in a broader sense - they characterize the expressive power of neural networks constructed simply by shifting and rescaling random features.

Information Laundering for Model Privacy

- Xinran Wang, Yu Xiang, Jun Gao, Jie Ding
- abstract@open-review(Spotlight): In this work, we propose information laundering, a novel framework for enhancing model privacy. Unlike data privacy that concerns the protection of raw data information, model privacy aims to protect an already-learned model that is to be deployed for public use. The private model can be obtained from general learning methods, and its deployment means that it will return a deterministic or random response for a given input query. An information-laundered model consists of probabilistic components that deliberately maneuver the intended input and output for queries of the model, so the model's adversarial acquisition is less likely. Under the proposed framework, we develop an information-theoretic principle to quantify the fundamental tradeoffs between model utility and privacy leakage and derive the optimal design.

Towards Faster and Stabilized GAN Training for High-fidelity Few-shot Image Synthesis

- Bingchen Liu, Yizhe Zhu, Kunpeng Song, Ahmed Elgammal
- abstract@open-review(Poster): Training Generative Adversarial Networks (GAN) on high-fidelity images usually requires large-scale GPU-clusters and a vast number of training images. In this paper, we study the few-shot image synthesis task for GAN with minimum computing cost. We propose a light-weight GAN structure that gains superior quality on 1024^2 resolution. Notably, the model converges from scratch with just a few hours of training on a single RTX-2080 GPU, and has a consistent performance, even with less than 100 training samples. Two technique designs constitute our work, a skip-layer channel-wise excitation module and a self-supervised discriminator trained as a feature-encoder. With thirteen datasets covering a wide variety of image domains (The datasets and code are available at https://github.com/odegeasslbc/FastGAN-pytorch), we show our model's superior performance compared to the state-of-the-art StyleGAN2, when data and computing budget are limited.

UPDeT: Universal Multi-agent RL via Policy Decoupling with Transformers

- Siyi Hu, Fengda Zhu, Xiaojun Chang, Xiaodan Liang
- abstract@open-review(Spotlight): Recent advances in multi-agent reinforcement learning have been largely limited in training one model from scratch for every new task. The limitation is due to the restricted model architecture related to fixed input and output dimensions. This hinders the experience accumulation and transfer of the learned agent over tasks with diverse levels of difficulty (e.g. 3 vs 3 or 5 vs 6 multi-agent games). In this paper, we make the first attempt to explore a universal multi-agent reinforcement learning pipeline, designing one single architecture to fit tasks with the requirement of different observation and action configurations. Unlike previous RNN-based models, we utilize a transformer-based model to generate a flexible policy by decoupling the policy distribution from the intertwined input observation with an importance weight measured by the merits of the self-attention mechanism. Compared to a standard transformer block, the proposed model, named as Universal Policy Decoupling Transformer (UPDeT), further relaxes the action restriction and makes the multi-agent task's decision process more explainable. UPDeT is general enough to be plugged into any multi-agent reinforcement learning pipeline and equip them with strong generalization abilities that enables the handling of multiple tasks at a time. Extensive experiments on large-scale SMAC multi-agent competitive games demonstrate that the proposed UPDeT-based multi-agent reinforcement learning achieves significant results relative to state-of-the-art approaches, demonstrating advantageous transfer capability in terms of both performance and training speed (10 times faster).

Free Lunch for Few-shot Learning: Distribution Calibration

- Shuo Yang, Lu Liu, Min Xu
- abstract@open-review(Oral): Learning from a limited number of samples is challenging since the learned model can easily become overfitted based on the biased distribution formed by only a few training examples. In this paper, we calibrate the distribution of these few-sample classes by transferring statistics from the classes with sufficient examples. Then an adequate number of examples can be sampled from the calibrated distribution to expand the inputs to the classifier. We assume every dimension in the feature representation follows a Gaussian distribution so that the mean and the variance of the distribution can borrow from that of similar classes whose statistics are better estimated with an adequate number of samples. Our method can be built on top of off-the-shelf pretrained feature extractors and classification models without extra parameters. We show that a simple logistic regression classifier trained using the features sampled from our calibrated distribution can outperform the state-of-the-art accuracy on three datasets (~5% improvement on miniImageNet compared to the next best). The visualization of these generated features demonstrates that our calibrated distribution is an accurate estimation.

Targeted Attack against Deep Neural Networks via Flipping Limited Weight Bits

- Jiawang Bai, Baoyuan Wu, Yong Zhang, Yiming Li, Zhifeng Li, Shu-Tao Xia
- abstract@open-review(Poster): To explore the vulnerability of deep neural networks (DNNs), many attack paradigms have been well studied, such as the poisoning-based backdoor attack in the training stage and the adversarial attack in the inference stage. In this paper, we study a novel attack paradigm, which modifies model parameters in the deployment stage for malicious purposes. Specifically, our goal is to misclassify a specific sample into a target class without any sample modification, while not significantly reduce the prediction accuracy of other samples to ensure the stealthiness. To this end, we formulate this problem as a binary integer programming (BIP), since the parameters are stored as binary bits ($i.e.$, 0 and 1) in the memory. By utilizing the latest technique in integer programming, we equivalently reformulate this BIP problem as a continuous optimization problem, which can be effectively and efficiently solved using the alternating direction method of multipliers (ADMM) method. Consequently, the flipped critical bits can be easily determined through optimization, rather than using a heuristic strategy. Extensive experiments demonstrate the superiority of our method in attacking DNNs.

FedBN: Federated Learning on Non-IID Features via Local Batch Normalization

- Xiaoxiao Li, Meirui JIANG, Xiaofei Zhang, Michael Kamp, Qi Dou
- abstract@open-review(Poster): The emerging paradigm of federated learning (FL) strives to enable collaborative training of deep models on the network edge without centrally aggregating raw data and hence improving data privacy. In most cases, the assumption of independent and identically distributed samples across local clients does not hold for federated learning setups. Under this setting, neural network training performance may vary significantly according to the data distribution and even hurt training convergence. Most of the previous work has focused on a difference in the distribution of labels or client shifts. Unlike those settings, we address an important problem of FL, e.g., different scanners/sensors in medical imaging, different scenery distribution in autonomous driving (highway vs. city), where local clients store examples with different distributions compared to other clients, which we denote as feature shift non-iid. In this work, we propose an effective method that uses local batch normalization to alleviate the feature shift before averaging models. The resulting scheme, called FedBN, outperforms both classical FedAvg, as well as the state-of-the-art for non-iid data (FedProx) on our extensive experiments. These empirical results are supported by a convergence analysis that shows in a simplified setting that FedBN has a faster convergence rate than FedAvg. Code is available at https://github.com/med-air/FedBN.

AdamP: Slowing Down the Slowdown for Momentum Optimizers on Scale-invariant Weights

- Byeongho Heo, Sanghyuk Chun, Seong Joon Oh, Dongyoon Han, Sangdoo Yun, Gyuwan Kim, Youngjung Uh, Jung-Woo Ha
- abstract@open-review(Poster): Normalization techniques, such as batch normalization (BN), are a boon for modern deep learning. They let weights converge more quickly with often better generalization performances. It has been argued that the normalization-induced scale invariance among the weights provides an advantageous ground for gradient descent (GD) optimizers: the effective step sizes are automatically reduced over time, stabilizing the overall training procedure. It is often overlooked, however, that the additional introduction of momentum in GD optimizers results in a far more rapid

reduction in effective step sizes for scale-invariant weights, a phenomenon that has not yet been studied and may have caused unwanted side effects in the current practice. This is a crucial issue because arguably the vast majority of modern deep neural networks consist of (1) momentum-based GD (e.g. SGD or Adam) and (2) scale-invariant parameters (e.g. more than 90% of the weights in ResNet are scale-invariant due to BN). In this paper, we verify that the widely-adopted combination of the two ingredients lead to the premature decay of effective step sizes and sub-optimal model performances. We propose a simple and effective remedy, SGDP and AdamP: get rid of the radial component, or the norm-increasing direction, at each optimizer step. Because of the scale invariance, this modification only alters the effective step sizes without changing the effective update directions, thus enjoying the original convergence properties of GD optimizers. Given the ubiquity of momentum GD and scale invariance in machine learning, we have evaluated our methods against the baselines on 13 benchmarks. They range from vision tasks like classification (e.g. ImageNet), retrieval (e.g. CUB and SOP), and detection (e.g. COCO) to language modelling (e.g. WikiText) and audio classification (e.g. DCASE) tasks. We verify that our solution brings about uniform gains in performances in those benchmarks. Source code is available at https://github.com/clovaai/adamp

Learning Subgoal Representations with Slow Dynamics

- Siyuan Li, Lulu Zheng, Jianhao Wang, Chongjie Zhang
- abstract@open-review(Poster): In goal-conditioned Hierarchical Reinforcement Learning (HRL), a high-level policy periodically sets subgoals for a low-level policy, and the low-level policy is trained to reach those subgoals. A proper subgoal representation function, which abstracts a state space to a latent subgoal space, is crucial for effective goal-conditioned HRL, since different low-level behaviors are induced by reaching subgoals in the compressed representation space. Observing that the high-level agent operates at an abstract temporal scale, we propose a slowness objective to effectively learn the subgoal representation (i.e., the high-level action space). We provide a theoretical grounding for the slowness objective. That is, selecting slow features as the subgoal space can achieve efficient hierarchical exploration. As a result of better exploration ability, our approach significantly outperforms state-of-the-art HRL and exploration methods on a number of benchmark continuous-control tasks. Thanks to the generality of the proposed subgoal representation learning method, empirical results also demonstrate that the learned representation and corresponding low-level policies can be transferred between distinct tasks.

A Unified Approach to Interpreting and Boosting Adversarial Transferability

- Xin Wang, Jie Ren, Shuyun Lin, Xiangming Zhu, Yisen Wang, Quanshi Zhang
- abstract@open-review(Poster): In this paper, we use the interaction inside adversarial perturbations to explain and boost the adversarial transferability. We discover and prove the negative correlation between the adversarial transferability and the interaction inside adversarial perturbations. The negative correlation is further verified through different DNNs with various inputs. Moreover, this negative correlation can be regarded as a unified perspective to understand current transferability-boosting methods. To this end, we prove that some classic methods of enhancing the transferability essentially decease interactions inside adversarial perturbations. Based on this, we propose to directly penalize interactions during the attacking process, which significantly improves the adversarial transferability. We will release the code when the paper is accepted.

Perceptual Adversarial Robustness: Defense Against Unseen Threat Models

- Cassidy Laidlaw, Sahil Singla, Soheil Feizi
- abstract@open-review(Poster): A key challenge in adversarial robustness is the lack of a precise mathematical characterization of human perception, used in the definition of adversarial attacks that are imperceptible to human eyes. Most current attacks and defenses try to get around this issue by considering restrictive adversarial threat models such as those bounded by L_2 or L_∞ distance, spatial perturbations, etc. However, models that are robust against any of these restrictive threat models are still fragile against other threat models, i.e. they have poor generalization to unforeseen attacks. Moreover, even if a model is robust against the union of several restrictive threat models, it is still susceptible to other imperceptible adversarial examples that are not contained in any of the constituent threat models. To resolve these issues, we propose adversarial training against the set of all imperceptible adversarial examples. Since this set is intractable to compute without a human in the loop, we approximate it using deep neural networks. We call this threat model the neural perceptual threat model (NPTM); it includes adversarial examples with a bounded neural perceptual distance (a neural network-based approximation of the true perceptual distance) to natural images. Through an extensive perceptual study, we show that the neural perceptual distance correlates well with human judgements of perceptibility of adversarial examples, validating our threat model.

Under the NPTM, we develop novel perceptual adversarial attacks and defenses. Because the NPTM is very broad, we find that Perceptual Adversarial Training (PAT) against a perceptual attack gives robustness against many other types of adversarial attacks. We test PAT on CIFAR-10 and ImageNet-100 against five diverse adversarial attacks: L_2, L_∞, spatial, recoloring, and JPEG. We find that PAT achieves state-of-the-art robustness against the union of these five attacks—more than doubling the accuracy over the next best model—without training against any of them. That is, PAT generalizes well to unforeseen perturbation types. This is vital in sensitive applications where a particular threat model cannot be assumed, and to the best of our knowledge, PAT is the first adversarial training defense with this property.

Code and data are available at https://github.com/cassidylaidlaw/perceptual-advex

Better Fine-Tuning by Reducing Representational Collapse

- Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, Sonal Gupta
- abstract@open-review(Poster): Although widely adopted, existing approaches for fine-tuning pre-trained language models have been shown to be unstable across hyper-parameter settings, motivating recent work on trust region methods. In this paper, we present a simplified and efficient method rooted in trust region theory that replaces previously used adversarial objectives with parametric noise (sampling from either a normal or uniform distribution), thereby discouraging representation change during fine-tuning when possible without hurting performance. We also introduce a new analysis to motivate the use of trust region methods more generally, by studying representational collapse; the degradation of generalizable representations from pre-trained models as they are fine-tuned for a specific end task. Extensive experiments show that our fine-tuning method matches or exceeds the performance of previous trust region methods on a range of understanding and generation tasks (including DailyMail/CNN, Gigaword, Reddit TIFU, and the GLUE benchmark), while also being much faster. We also show that it is less prone to representation collapse; the pre-trained models maintain more generalizable representations every time they are fine-tuned.

Learning N:M Fine-grained Structured Sparse Neural Networks From Scratch

- Aojun Zhou, Yukun Ma, Junnan Zhu, Jianbo Liu, Zhijie Zhang, Kun Yuan, Wenxiu Sun, Hongsheng Li
- abstract@open-review(Poster): Sparsity in Deep Neural Networks (DNNs) has been widely studied to compress and accelerate the models on resource-constrained environments. It can be generally categorized into unstructured fine-grained sparsity that zeroes out multiple individual weights distributed across the neural network, and structured coarse-grained sparsity which prunes blocks of sub-networks of a neural network. Fine-grained sparsity can achieve a high compression ratio but is not hardware friendly and hence receives limited speed gains. On the other hand, coarse-grained sparsity cannot simultaneously achieve both apparent acceleration on modern GPUs and decent performance. In this paper, we are the first to study training from scratch an N:M fine-grained structured sparse network, which can maintain the advantages of both unstructured fine-grained sparsity and structured coarse-grained sparsity simultaneously on specifically designed GPUs. Specifically, a 2 : 4 sparse network could achieve 2× speed-up without performance drop on Nvidia A100 GPUs. Furthermore, we propose a novel and effective ingredient, sparse-refined straight-through estimator (SR-STE), to alleviate the negative influence of the approximated gradients computed by vanilla STE during optimization. We also define a metric, Sparse Architecture Divergence

(SAD), to measure the sparse network's topology change during the training process. Finally, We justify SR-STE's advantages with SAD and demonstrate the effectiveness of SR-STE by performing comprehensive experiments on various tasks. Anonymous code and model will be at available at https://github.com/anonymous-NM-sparsity/NM-sparsity.

Active Contrastive Learning of Audio-Visual Video Representations

- Shuang Ma, Zhaoyang Zeng, Daniel McDuff, Yale Song
- abstract@open-review(Poster): Contrastive learning has been shown to produce generalizable representations of audio and visual data by maximizing the lower bound on the mutual information (MI) between different views of an instance. However, obtaining a tight lower bound requires a sample size exponential in MI and thus a large set of negative samples. We can incorporate more samples by building a large queue-based dictionary, but there are theoretical limits to performance improvements even with a large number of negative samples. We hypothesize that random negative sampling leads to a highly redundant dictionary that results in suboptimal representations for downstream tasks. In this paper, we propose an active contrastive learning approach that builds an actively sampled dictionary with diverse and informative items, which improves the quality of negative samples and improves performances on tasks where there is high mutual information in the data, e.g., video classification. Our model achieves state-of-the-art performance on challenging audio and visual downstream benchmarks including UCF101, HMDB51 and ESC50.

PseudoSeg: Designing Pseudo Labels for Semantic Segmentation

- Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, Tomas Pfister
- abstract@open-review(Poster): Recent advances in semi-supervised learning (SSL) demonstrate that a combination of consistency regularization and pseudo-labeling can effectively improve image classification accuracy in the low-data regime. Compared to classification, semantic segmentation tasks require much more intensive labeling costs. Thus, these tasks greatly benefit from data-efficient training methods. However, structured outputs in segmentation render particular difficulties (e.g., designing pseudo-labeling and augmentation) to apply existing SSL strategies. To address this problem, we present a simple and novel re-design of pseudo-labeling to generate well-calibrated structured pseudo labels for training with unlabeled or weakly-labeled data. Our proposed pseudo-labeling strategy is network structure agnostic to apply in a one-stage consistency training framework. We demonstrate the effectiveness of the proposed pseudo-labeling strategy in both low-data and high-data regimes. Extensive experiments have validated that pseudo labels generated from wisely fusing diverse sources and strong data augmentation are crucial to consistency training for segmentation. The source code will be released.

Correcting experience replay for multi-agent communication

- Sanjeevan Ahilan, Peter Dayan
- abstract@open-review(Spotlight): We consider the problem of learning to communicate using multi-agent reinforcement learning (MARL). A common approach is to learn off-policy, using data sampled from a replay buffer. However, messages received in the past may not accurately reflect the current communication policy of each agent, and this complicates learning. We therefore introduce a 'communication correction' which accounts for the non-stationarity of observed communication induced by multi-agent learning. It works by relabelling the received message to make it likely under the communicator's current policy, and thus be a better reflection of the receiver's current environment. To account for cases in which agents are both senders and receivers, we introduce an ordered relabelling scheme. Our correction is computationally efficient and can be integrated with a range of off-policy algorithms. We find in our experiments that it substantially improves the ability of communicating MARL systems to learn across a variety of cooperative and competitive tasks.

Counterfactual Generative Networks

- Axel Sauer, Andreas Geiger
- abstract@open-review(Poster): Neural networks are prone to learning shortcuts -- they often model simple correlations, ignoring more complex ones that potentially generalize better. Prior works on image classification show that instead of learning a connection to object shape, deep classifiers tend to exploit spurious correlations with low-level texture or the background for solving the classification task. In this work, we take a step towards more robust and interpretable classifiers that explicitly expose the task's causal structure. Building on current advances in deep generative modeling, we propose to decompose the image generation process into independent causal mechanisms that we train without direct supervision. By exploiting appropriate inductive biases, these mechanisms disentangle object shape, object texture, and background; hence, they allow for generating counterfactual images. We demonstrate the ability of our model to generate such images on MNIST and ImageNet. Further, we show that the counterfactual images can improve out-of-distribution robustness with a marginal drop in performance on the original classification task, despite being synthetic. Lastly, our generative model can be trained efficiently on a single GPU, exploiting common pre-trained models as inductive biases.

Contemplating Real-World Object Classification

- ali borji
- abstract@open-review(Poster): Deep object recognition models have been very successful over benchmark datasets such as ImageNet. How accurate and robust are they to distribution shifts arising from natural and synthetic variations in datasets? Prior research on this problem has primarily focused on ImageNet variations (e.g., ImageNetV2, ImageNet-A). To avoid potential inherited biases in these studies, we take a different approach. Specifically, we reanalyze the ObjectNet dataset recently proposed by Barbu et al. containing objects in daily life situations. They showed a dramatic performance drop of the state of the art object recognition models on this dataset. Due to the importance and implications of their results regarding the generalization ability of deep models, we take a second look at their analysis. We find that applying deep models to the isolated objects, rather than the entire scene as is done in the original paper, results in around 20-30% performance improvement. Relative to the numbers reported in Barbu et al., around 10-15% of the performance loss is recovered, without any test time data augmentation. Despite this gain, however, we conclude that deep models still suffer drastically on the ObjectNet dataset. We also investigate the robustness of models against synthetic image perturbations such as geometric transformations (e.g., scale, rotation, translation), natural image distortions (e.g., impulse noise, blur) as well as adversarial attacks (e.g., FGSM and PGD-5). Our results indicate that limiting the object area as much as possible (i.e., from the entire image to the bounding box to the segmentation mask) leads to consistent improvement in accuracy and robustness. Finally, through a qualitative analysis of ObjectNet data, we find that i) a large number of images in this dataset are hard to recognize even for humans, and ii) easy (hard) samples for models match with easy (hard) samples for humans. Overall, our analysis shows that ObjecNet is still a challenging test platform that can be used to measure the generalization ability of models. The code and data are available in [masked due to blind review].

Towards Nonlinear Disentanglement in Natural Data with Temporal Sparse Coding

- David A. Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, Dylan Paiton
- abstract@open-review(Oral): Disentangling the underlying generative factors from complex data has so far been limited to carefully constructed scenarios. We propose a path towards natural data by first showing that the statistics of natural data provide enough structure to enable disentanglement, both theoretically and empirically. Specifically, we provide evidence that objects in natural movies undergo transitions that are typically small in magnitude with occasional large jumps, which is characteristic of a temporally sparse distribution. To address this finding we provide a novel proof that relies on a sparse prior on temporally adjacent observations to recover the true latent variables up to permutations and sign flips, directly providing a stronger result than previous work. We show that equipping practical estimation methods with our prior often surpasses the current state-of-the-art on

several established benchmark datasets without any impractical assumptions, such as knowledge of the number of changing generative factors. Furthermore, we contribute two new benchmarks, Natural Sprites and KITTI Masks, which integrate the measured natural dynamics to enable disentanglement evaluation with more realistic datasets. We leverage these benchmarks to test our theory, demonstrating improved performance. We also identify non-obvious challenges for current methods in scaling to more natural domains. Taken together our work addresses key issues in disentanglement research for moving towards more natural settings.

Explainable Deep One-Class Classification

- Philipp Liznerski, Lukas Ruff, Robert A. Vandermeulen, Billy Joe Franks, Marius Kloft, Klaus Robert Muller
- abstract@open-review(Poster): Deep one-class classification variants for anomaly detection learn a mapping that concentrates nominal samples in feature space causing anomalies to be mapped away. Because this transformation is highly non-linear, finding interpretations poses a significant challenge. In this paper we present an explainable deep one-class classification method, Fully Convolutional Data Description (FCDD), where the mapped samples are themselves also an explanation heatmap. FCDD yields competitive detection performance and provides reasonable explanations on common anomaly detection benchmarks with CIFAR-10 and ImageNet. On MVTec-AD, a recent manufacturing dataset offering ground-truth anomaly maps, FCDD sets a new state of the art in the unsupervised setting. Our method can incorporate ground-truth anomaly maps during training and using even a few of these (~5) improves performance significantly. Finally, using FCDD's explanations we demonstrate the vulnerability of deep one-class classification models to spurious image features such as image watermarks.

Batch Reinforcement Learning Through Continuation Method

- Yijie Guo, Shengyu Feng, Nicolas Le Roux, Ed Chi, Honglak Lee, Minmin Chen
- abstract@open-review(Poster): Many real-world applications of reinforcement learning (RL) require the agent to learn from a fixed set of trajectories, without collecting new interactions. Policy optimization under this setting is extremely challenging as: 1) the geometry of the objective function is hard to optimize efficiently; 2) the shift of data distributions causes high noise in the value estimation. In this work, we propose a simple yet effective policy iteration approach to batch RL using global optimization techniques known as continuation. By constraining the difference between the learned policy and the behavior policy that generates the fixed trajectories, and continuously relaxing the constraint, our method 1) helps the agent escape local optima; 2) reduces the error in policy evaluation in the optimization procedure. We present results on a variety of control tasks, game environments, and a recommendation task to empirically demonstrate the efficacy of our proposed method.

Protecting DNNs from Theft using an Ensemble of Diverse Models

- Sanjay Kariyappa, Atul Prakash, Moinuddin K Qureshi
- abstract@open-review(Poster): Several recent works have demonstrated highly effective model stealing (MS) attacks on Deep Neural Networks (DNNs) in black-box settings, even when the training data is unavailable. These attacks typically use some form of Out of Distribution (OOD) data to query the target model and use the predictions obtained to train a clone model. Such a clone model learns to approximate the decision boundary of the target model, achieving high accuracy on in-distribution examples. We propose Ensemble of Diverse Models (EDM) to defend against such MS attacks. EDM is made up of models that are trained to produce dissimilar predictions for OOD inputs. By using a different member of the ensemble to service different queries, our defense produces predictions that are highly discontinuous in the input space for the adversary's OOD queries. Such discontinuities cause the clone model trained on these predictions to have poor generalization on in-distribution examples. Our evaluations on several image classification tasks demonstrate that EDM defense can severely degrade the accuracy of clone models (up to 39.7%). Our defense has minimal impact on the target accuracy, negligible computational costs during inference, and is compatible with existing defenses for MS attacks.

Domain Generalization with MixStyle

- Kaiyang Zhou, Yongxin Yang, Yu Qiao, Tao Xiang
- abstract@open-review(Poster): Though convolutional neural networks (CNNs) have demonstrated remarkable ability in learning discriminative features, they often generalize poorly to unseen domains. Domain generalization aims to address this problem by learning from a set of source domains a model that is generalizable to any unseen domain. In this paper, a novel approach is proposed based on probabilistically mixing instance-level feature statistics of training samples across source domains. Our method, termed MixStyle, is motivated by the observation that visual domain is closely related to image style (e.g., photo vs.~sketch images). Such style information is captured by the bottom layers of a CNN where our proposed style-mixing takes place. Mixing styles of training instances results in novel domains being synthesized implicitly, which increase the domain diversity of the source domains, and hence the generalizability of the trained model. MixStyle fits into mini-batch training perfectly and is extremely easy to implement. The effectiveness of MixStyle is demonstrated on a wide range of tasks including category classification, instance retrieval and reinforcement learning.

Efficient Inference of Flexible Interaction in Spiking-neuron Networks

- Feng Zhou, Yixuan Zhang, Jun Zhu
- abstract@open-review(Poster): Hawkes process provides an effective statistical framework for analyzing the time-dependent interaction of neuronal spiking activities. Although utilized in many real applications, the classic Hawkes process is incapable of modelling inhibitory interactions among neurons. Instead, the nonlinear Hawkes process allows for a more flexible influence pattern with excitatory or inhibitory interactions. In this paper, three sets of auxiliary latent variables (Polya-Gamma variables, latent marked Poisson processes and sparsity variables) are augmented to make functional connection weights in a Gaussian form, which allows for a simple iterative algorithm with analytical updates. As a result, an efficient expectation-maximization (EM) algorithm is derived to obtain the maximum a posteriori (MAP) estimate. We demonstrate the accuracy and efficiency performance of our algorithm on synthetic and real data. For real neural recordings, we show our algorithm can estimate the temporal dynamics of interaction and reveal the interpretable functional connectivity underlying neural spike trains.

DICE: Diversity in Deep Ensembles via Conditional Redundancy Adversarial Estimation

- Alexandre Rame, Matthieu Cord
- abstract@open-review(Poster): Deep ensembles perform better than a single network thanks to the diversity among their members. Recent approaches regularize predictions to increase diversity; however, they also drastically decrease individual members' performances. In this paper, we argue that learning strategies for deep ensembles need to tackle the trade-off between ensemble diversity and individual accuracies. Motivated by arguments from information theory and leveraging recent advances in neural estimation of conditional mutual information, we introduce a novel training criterion called DICE: it increases diversity by reducing spurious correlations among features. The main idea is that features extracted from pairs of members should only share information useful for target class prediction without being conditionally redundant. Therefore, besides the classification loss with information bottleneck, we adversarially prevent features from being conditionally predictable from each other. We manage to reduce simultaneous errors while protecting class information. We obtain state-of-the-art accuracy results on CIFAR-10/100: for example, an ensemble of 5 networks trained with DICE matches an ensemble of 7 networks trained independently. We further analyze the consequences on calibration, uncertainty estimation, out-of-distribution detection and online co-distillation.

Universal Weakly Supervised Segmentation by Pixel-to-Segment Contrastive Learning

- Tsung-Wei Ke, Jyh-Jing Hwang, Stella Yu
- abstract@open-review(Poster): Weakly supervised segmentation requires assigning a label to every pixel based on training instances with partial annotations such as image-level tags, object bounding boxes, labeled points and scribbles. This task is challenging, as coarse annotations (tags, boxes) lack precise pixel localization whereas sparse annotations (points, scribbles) lack broad region coverage. Existing methods tackle these two types of weak supervision differently: Class activation maps are used to localize coarse labels and iteratively refine the segmentation model, whereas conditional random fields are used to propagate sparse labels to the entire image.

We formulate weakly supervised segmentation as a semi-supervised metric learning problem, where pixels of the same (different) semantics need to be mapped to the same (distinctive) features. We propose 4 types of contrastive relationships between pixels and segments in the feature space, capturing low-level image similarity, semantic annotation, co-occurrence, and feature affinity They act as priors; the pixel-wise feature can be learned from training images with any partial annotations in a data-driven fashion. In particular, unlabeled pixels in training images participate not only in data-driven grouping within each image, but also in discriminative feature learning within and across images. We deliver a universal weakly supervised segmenter with significant gains on Pascal VOC and DensePose. Our code is publicly available at https://github.com/twke18/SPML.

Shape-Texture Debiased Neural Network Training

- Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan Yuille, cihang xie
- abstract@open-review(Poster): Shape and texture are two prominent and complementary cues for recognizing objects. Nonetheless, Convolutional Neural Networks are often biased towards either texture or shape, depending on the training dataset. Our ablation shows that such bias degenerates model performance. Motivated by this observation, we develop a simple algorithm for shape-texture debiased learning. To prevent models from exclusively attending on a single cue in representation learning, we augment training data with images with conflicting shape and texture information (eg, an image of chimpanzee shape but with lemon texture) and, most importantly, provide the corresponding supervisions from shape and texture simultaneously.

Experiments show that our method successfully improves model performance on several image recognition benchmarks and adversarial robustness. For example, by training on ImageNet, it helps ResNet-152 achieve substantial improvements on ImageNet (+1.2%), ImageNet-A (+5.2%), ImageNet-C (+8.3%) and Stylized-ImageNet (+11.1%), and on defending against FGSM adversarial attacker on ImageNet (+14.4%). Our method also claims to be compatible with other advanced data augmentation strategies, eg, Mixup, and CutMix. The code is available here: https://github.com/LiYingwei/ShapeTextureDebiasedTraining.

Interactive Weak Supervision: Learning Useful Heuristics for Data Labeling

- Benedikt Boecking, Willie Neiswanger, Eric Xing, Artur Dubrawski
- abstract@open-review(Poster): Obtaining large annotated datasets is critical for training successful machine learning models and it is often a bottleneck in practice. Weak supervision offers a promising alternative for producing labeled datasets without ground truth annotations by generating probabilistic labels using multiple noisy heuristics. This process can scale to large datasets and has demonstrated state of the art performance in diverse domains such as healthcare and e-commerce. One practical issue with learning from user-generated heuristics is that their creation requires creativity, foresight, and domain expertise from those who hand-craft them, a process which can be tedious and subjective. We develop the first framework for interactive weak supervision in which a method proposes heuristics and learns from user feedback given on each proposed heuristic. Our experiments demonstrate that only a small number of feedback iterations are needed to train models that achieve highly competitive test set performance without access to ground truth training labels. We conduct user studies, which show that users are able to effectively provide feedback on heuristics and that test set results track the performance of simulated oracles.

MoVie: Revisiting Modulated Convolutions for Visual Counting and Beyond

- Duy Kien Nguyen, Vedanuj Goswami, Xinlei Chen
- abstract@open-review(Poster): This paper focuses on visual counting, which aims to predict the number of occurrences given a natural image and a query (e.g. a question or a category). Unlike most prior works that use explicit, symbolic models which can be computationally expensive and limited in generalization, we propose a simple and effective alternative by revisiting modulated convolutions that fuse the query and the image locally. Following the design of residual bottleneck, we call our method MoVie, short for Modulated conVolutional bottlenecks. Notably, MoVie reasons implicitly and holistically and only needs a single forward-pass during inference. Nevertheless, MoVie showcases strong performance for counting: 1) advancing the state-of-the-art on counting-specific VQA tasks while being more efficient; 2) outperforming prior-art on difficult benchmarks like COCO for common object counting; 3) helped us secure the first place of 2020 VQA challenge when integrated as a module for ‘number’ related questions in generic VQA models. Finally, we show evidence that modulated convolutions such as MoVie can serve as a general mechanism for reasoning tasks beyond counting.

Improve Object Detection with Feature-based Knowledge Distillation: Towards Accurate and Efficient Detectors

- Linfeng Zhang, Kaisheng Ma
- abstract@open-review(Poster): Knowledge distillation, in which a student model is trained to mimic a teacher model, has been proved as an effective technique for model compression and model accuracy boosting. However, most knowledge distillation methods, designed for image classification, have failed on more challenging tasks, such as object detection. In this paper, we suggest that the failure of knowledge distillation on object detection is mainly caused by two reasons: (1) the imbalance between pixels of foreground and background and (2) lack of distillation on the relation between different pixels. Observing the above reasons, we propose attention-guided distillation and non-local distillation to address the two problems, respectively. Attention-guided distillation is proposed to find the crucial pixels of foreground objects with attention mechanism and then make the students take more effort to learn their features. Non-local distillation is proposed to enable students to learn not only the feature of an individual pixel but also the relation between different pixels captured by non-local modules. Experiments show that our methods achieve excellent AP improvements on both one-stage and two-stage, both anchor-based and anchor-free detectors. For example, Faster RCNN (ResNet101 backbone) with our distillation achieves 43.9 AP on COCO2017, which is 4.1 higher than the baseline. Codes have been released on Github.

Revisiting Dynamic Convolution via Matrix Decomposition

- Yunsheng Li, Yinpeng Chen, Xiyang Dai, mengchen liu, Dongdong Chen, Ye Yu, Lu Yuan, Zicheng Liu, Mei Chen, Nuno Vasconcelos
- abstract@open-review(Poster): Recent research in dynamic convolution shows substantial performance boost for efficient CNNs, due to the adaptive aggregation of K static convolution kernels. It has two limitations: (a) it increases the number of convolutional weights by K-times, and (b) the joint optimization of dynamic attention and static convolution kernels is challenging. In this paper, we revisit it from a new perspective of matrix decomposition and reveal the key issue is that dynamic convolution applies dynamic attention over channel groups after projecting into a higher dimensional latent space. To address this issue, we propose dynamic channel fusion to replace dynamic attention over channel groups. Dynamic channel fusion not only enables significant dimension reduction of the latent space, but also mitigates the joint optimization difficulty. As a result, our method is easier to train and requires significantly fewer parameters without sacrificing accuracy. Source code is at https://github.com/liyunsheng13/dcd.

Improving Adversarial Robustness via Channel-wise Activation Suppressing

- Yang Bai, Yuyuan Zeng, Yong Jiang, Shu-Tao Xia, Xingjun Ma, Yisen Wang

- abstract@open-review(Spotlight): The study of adversarial examples and their activations have attracted significant attention for secure and robust learning with deep neural networks (DNNs). Different from existing works, in this paper, we highlight two new characteristics of adversarial examples from the channel-wise activation perspective: 1) the activation magnitudes of adversarial examples are higher than that of natural examples; and 2) the channels are activated more uniformly by adversarial examples than natural examples. We find that, while the state-of-the-art defense adversarial training has addressed the first issue of high activation magnitude via training on adversarial examples, the second issue of uniform activation remains. This motivates us to suppress redundant activations from being activated by adversarial perturbations during the adversarial training process, via a Channel-wise Activation Suppressing (CAS) training strategy. We show that CAS can train a model that inherently suppresses adversarial activations, and can be easily applied to existing defense methods to further improve their robustness. Our work provides a simplebut generic training strategy for robustifying the intermediate layer activations of DNNs.

DynaTune: Dynamic Tensor Program Optimization in Deep Neural Network Compilation

- Minjia Zhang, Menghao Li, Chi Wang, Mingqin Li
- abstract@open-review(Poster): Recently, the DL compiler, together with Learning to Compile has proven to be a powerful technique for optimizing deep learning models. However, existing methods focus on accelerating the convergence speed of the individual tensor operator rather than the convergence speed of the entire model, which results in long optimization time to obtain a desired latency.

In this paper, we present a new method called DynaTune, which provides significantly faster convergence speed to optimize a DNN model. In particular, we consider a Multi-Armed Bandit (MAB) model for the tensor program optimization problem. We use UCB to handle the decision-making of time-slot-based optimization, and we devise a Bayesian belief model that allows predicting the potential performance gain of each operator with uncertainty quantification, which guides the optimization process. We evaluate and compare DynaTune with the state-of-the-art DL compiler. The experiment results show that DynaTune is 1.2--2.4 times faster to achieve the same optimization quality for a range of models across different hardware architectures.

MALI: A memory efficient and reverse accurate integrator for Neural ODEs

- Juntang Zhuang, Nicha C Dvornek, sekhar tatikonda, James s Duncan
- abstract@open-review(Poster): Neural ordinary differential equations (Neural ODEs) are a new family of deep-learning models with continuous depth. However, the numerical estimation of the gradient in the continuous case is not well solved: existing implementations of the adjoint method suffer from inaccuracy in reverse-time trajectory, while the naive method and the adaptive checkpoint adjoint method (ACA) have a memory cost that grows with integration time. In this project, based on the asynchronous leapfrog (ALF) solver, we propose the Memory-efficient ALF Integrator (MALI), which has a constant memory cost $w.r.t$ integration time similar to the adjoint method, and guarantees accuracy in reverse-time trajectory (hence accuracy in gradient estimation). We validate MALI in various tasks: on image recognition tasks, to our knowledge, MALI is the first to enable feasible training of a Neural ODE on ImageNet and outperform a well-tuned ResNet, while existing methods fail due to either heavy memory burden or inaccuracy; for time series modeling, MALI significantly outperforms the adjoint method; and for continuous generative models, MALI achieves new state-of-the-art performance. We provide a pypi package: https://jzkay12.github.io/TorchDiffEqPack

Bag of Tricks for Adversarial Training

- Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, Jun Zhu
- abstract@open-review(Poster): Adversarial training (AT) is one of the most effective strategies for promoting model robustness. However, recent benchmarks show that most of the proposed improvements on AT are less effective than simply early stopping the training procedure. This counter-intuitive fact motivates us to investigate the implementation details of tens of AT methods. Surprisingly, we find that the basic settings (e.g., weight decay, training schedule, etc.) used in these methods are highly inconsistent. In this work, we provide comprehensive evaluations on CIFAR-10, focusing on the effects of mostly overlooked training tricks and hyperparameters for adversarially trained models. Our empirical observations suggest that adversarial robustness is much more sensitive to some basic training settings than we thought. For example, a slightly different value of weight decay can reduce the model robust accuracy by more than 7%, which is probable to override the potential promotion induced by the proposed methods. We conclude a baseline training setting and re-implement previous defenses to achieve new state-of-the-art results. These facts also appeal to more concerns on the overlooked confounders when benchmarking defenses.

Learning to Generate 3D Shapes with Generative Cellular Automata

- Dongsu Zhang, Changwoon Choi, Jeonghwan Kim, Young Min Kim
- abstract@open-review(Poster): In this work, we present a probabilistic 3D generative model, named Generative Cellular Automata, which is able to produce diverse and high quality shapes. We formulate the shape generation process as sampling from the transition kernel of a Markov chain, where the sampling chain eventually evolves to the full shape of the learned distribution. The transition kernel employs the local update rules of cellular automata, effectively reducing the search space in a high-resolution 3D grid space by exploiting the connectivity and sparsity of 3D shapes. Our progressive generation only focuses on the sparse set of occupied voxels and their neighborhood, thus enables the utilization of an expressive sparse convolutional network. We propose an effective training scheme to obtain the local homogeneous rule of generative cellular automata with sequences that are slightly different from the sampling chain but converge to the full shapes in the training data. Extensive experiments on probabilistic shape completion and shape generation demonstrate that our method achieves competitive performance against recent methods.

SEED: Self-supervised Distillation For Visual Representation

- Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, Zicheng Liu
- abstract@open-review(Poster): This paper is concerned with self-supervised learning for small models. The problem is motivated by our empirical studies that while the widely used contrastive self-supervised learning method has shown great progress on large model training, it does not work well for small models. To address this problem, we propose a new learning paradigm, named $\textbf{SE}$lf-Sup$\textbf{E}$rvised $\textbf{D}$istillation (${\large S}$EED), where we leverage a larger network (as Teacher) to transfer its representational knowledge into a smaller architecture (as Student) in a self-supervised fashion. Instead of directly learning from unlabeled data, we train a student encoder to mimic the similarity score distribution inferred by a teacher over a set of instances. We show that ${\large S}$EED dramatically boosts the performance of small networks on downstream tasks. Compared with self-supervised baselines, ${\large S}$EED improves the top-1 accuracy from 42.2% to 67.6% on EfficientNet-B0 and from 36.3% to 68.2% on MobileNet-v3-Large on the ImageNet-1k dataset.

Long-tailed Recognition by Routing Diverse Distribution-Aware Experts

- Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, Stella Yu
- abstract@open-review(Spotlight): Natural data are often long-tail distributed over semantic classes. Existing recognition methods tackle this imbalanced classification by placing more emphasis on the tail data, through class re-balancing/re-weighting or ensembling over different data groups, resulting in increased tail accuracies but reduced head accuracies. We take a dynamic view of the training data and provide a principled model bias and variance analysis as the training data fluctuates: Existing long-tail classifiers invariably increase the model variance and the head-tail model bias gap remains large, due to more and larger confusion with hard negatives for the tail. We propose a new long-tailed classifier called RoutIng Diverse Experts (RIDE). It reduces the model variance with multiple experts, reduces the model bias with a distribution-aware diversity loss, reduces the computational cost with a

dynamic expert routing module. RIDE outperforms the state-of-the-art by 5% to 7% on CIFAR100-LT, ImageNet-LT and iNaturalist 2018 benchmarks. It is also a universal framework that is applicable to various backbone networks, long-tailed algorithms and training mechanisms for consistent performance gains. Our code is available at: https://github.com/frank-xwang/RIDE-LongTailRecognition.

PDE-Driven Spatiotemporal Disentanglement

- Jérémie Donà, Jean-Yves Franceschi, sylvain lamprier, patrick gallinari
- abstract@open-review(Poster): A recent line of work in the machine learning community addresses the problem of predicting high-dimensional spatiotemporal phenomena by leveraging specific tools from the differential equations theory. Following this direction, we propose in this article a novel and general paradigm for this task based on a resolution method for partial differential equations: the separation of variables. This inspiration allows us to introduce a dynamical interpretation of spatiotemporal disentanglement. It induces a principled model based on learning disentangled spatial and temporal representations of a phenomenon to accurately predict future observations. We experimentally demonstrate the performance and broad applicability of our method against prior state-of-the-art models on physical and synthetic video datasets.

Generalization in data-driven models of primary visual cortex

- Konstantin-Klemens Lurz, Mohammad Bashiri, Konstantin Willeke, Akshay Jagadish, Eric Wang, Edgar Y. Walker, Santiago A Cadena, Taliah Muhammad, Erick Cobos, Andreas S. Tolias, Alexander S Ecker, Fabian H. Sinz
- abstract@open-review(Spotlight): Deep neural networks (DNN) have set new standards at predicting responses of neural populations to visual input. Most such DNNs consist of a convolutional network (core) shared across all neurons which learns a representation of neural computation in visual cortex and a neuron-specific readout that linearly combines the relevant features in this representation. The goal of this paper is to test whether such a representation is indeed generally characteristic for visual cortex, i.e. generalizes between animals of a species, and what factors contribute to obtaining such a generalizing core. To push all non-linear computations into the core where the generalizing cortical features should be learned, we devise a novel readout that reduces the number of parameters per neuron in the readout by up to two orders of magnitude compared to the previous state-of-the-art. It does so by taking advantage of retinotopy and learns a Gaussian distribution over the neuron's receptive field position. With this new readout we train our network on neural responses from mouse primary visual cortex (V1) and obtain a gain in performance of 7% compared to the previous state-of-the-art network. We then investigate whether the convolutional core indeed captures general cortical features by using the core in transfer learning to a different animal. When transferring a core trained on thousands of neurons from various animals and scans we exceed the performance of training directly on that animal by 12%, and outperform a commonly used VGG16 core pre-trained on imagenet by 33%. In addition, transfer learning with our data-driven core is more data-efficient than direct training, achieving the same performance with only 40% of the data. Our model with its novel readout thus sets a new state-of-the-art for neural response prediction in mouse visual cortex from natural images, generalizes between animals, and captures better characteristic cortical features than current task-driven pre-training approaches such as VGG16.

Do 2D GANs Know 3D Shape? Unsupervised 3D Shape Reconstruction from 2D Image GANs

- Xingang Pan, Bo Dai, Ziwei Liu, Chen Change Loy, Ping Luo
- abstract@open-review(Oral): Natural images are projections of 3D objects on a 2D image plane. While state-of-the-art 2D generative models like GANs show unprecedented quality in modeling the natural image manifold, it is unclear whether they implicitly capture the underlying 3D object structures. And if so, how could we exploit such knowledge to recover the 3D shapes of objects in the images? To answer these questions, in this work, we present the first attempt to directly mine 3D geometric cues from an off-the-shelf 2D GAN that is trained on RGB images only. Through our investigation, we found that such a pre-trained GAN indeed contains rich 3D knowledge and thus can be used to recover 3D shape from a single 2D image in an unsupervised manner. The core of our framework is an iterative strategy that explores and exploits diverse viewpoint and lighting variations in the GAN image manifold. The framework does not require 2D keypoint or 3D annotations, or strong assumptions on object shapes (e.g. shapes are symmetric), yet it successfully recovers 3D shapes with high precision for human faces, cats, cars, and buildings. The recovered 3D shapes immediately allow high-quality image editing like relighting and object rotation. We quantitatively demonstrate the effectiveness of our approach compared to previous methods in both 3D shape reconstruction and face rotation. Our code is available at https://github.com/XingangPan/GAN2Shape.

Rethinking Soft Labels for Knowledge Distillation: A Bias–Variance Tradeoff Perspective

- Helong Zhou, Liangchen Song, Jiajie Chen, Ye Zhou, Guoli Wang, Junsong Yuan, Qian Zhang
- abstract@open-review(Poster): Knowledge distillation is an effective approach to leverage a well-trained network or an ensemble of them, named as the teacher, to guide the training of a student network. The outputs from the teacher network are used as soft labels for supervising the training of a new network. Recent studies (M üller et al., 2019; Yuan et al., 2020) revealed an intriguing property of the soft labels that making labels soft serves as a good regularization to the student network. From the perspective of statistical learning, regularization aims to reduce the variance, however how bias and variance change is not clear for training with soft labels. In this paper, we investigate the bias-variance tradeoff brought by distillation with soft labels. Specifically, we observe that during training the bias-variance tradeoff varies sample-wisely. Further, under the same distillation temperature setting, we observe that the distillation performance is negatively associated with the number of some specific samples, which are named as regularization samples since these samples lead to bias increasing and variance decreasing. Nevertheless, we empirically find that completely filtering out regularization samples also deteriorates distillation performance. Our discoveries inspired us to propose the novel weighted soft labels to help the network adaptively handle the sample-wise bias-variance tradeoff. Experiments on standard evaluation benchmarks validate the effectiveness of our method. Our code is available in the supplementary.

Distributed Momentum for Byzantine-resilient Stochastic Gradient Descent

- El Mahdi El Mhamdi, Rachid Guerraoui, Sébastien Rouault
- abstract@open-review(Poster): Byzantine-resilient Stochastic Gradient Descent (SGD) aims at shielding model training from Byzantine faults, be they ill-labeled training datapoints, exploited software/hardware vulnerabilities, or malicious worker nodes in a distributed setting. Two recent attacks have been challenging state-of-the-art defenses though, often successfully precluding the model from even fitting the training set. The main identified weakness in current defenses is their requirement of a sufficiently low variance-norm ratio for the stochastic gradients. We propose a practical method which, despite increasing the variance, reduces the variance-norm ratio, mitigating the identified weakness. We assess the effectiveness of our method over 736 different training configurations, comprising the 2 state-of-the-art attacks and 6 defenses. For confidence and reproducibility purposes, each configuration is run 5 times with specified seeds (1 to 5), totalling 3680 runs. In our experiments, when the attack is effective enough to decrease the highest observed top-1 cross-accuracy by at least 20% compared to the unattacked run, our technique systematically increases back the highest observed accuracy, and is able to recover at least 20% in more than 60% of the cases.

Scaling the Convex Barrier with Active Sets

- Alessandro De Palma, Harkirat Behl, Rudy R Bunel, Philip Torr, M. Pawan Kumar
- abstract@open-review(Poster): Tight and efficient neural network bounding is of critical importance for the scaling of neural network verification systems. A number of efficient specialised dual solvers for neural network bounds have been presented recently, but they are often too loose to verify more challenging properties. This lack of tightness is linked to the weakness of the employed relaxation, which is usually a linear program of size linear in the number of neurons. While a tighter linear relaxation for piecewise linear activations exists, it comes at the cost of exponentially many constraints and thus

currently lacks an efficient customised solver. We alleviate this deficiency via a novel dual algorithm that realises the full potential of the new relaxation by operating on a small active set of dual variables. Our method recovers the strengths of the new relaxation in the dual space: tightness and a linear separation oracle. At the same time, it shares the benefits of previous dual approaches for weaker relaxations: massive parallelism, GPU implementation, low cost per iteration and valid bounds at any time. As a consequence, we obtain better bounds than off-the-shelf solvers in only a fraction of their running time and recover the speed-accuracy trade-offs of looser dual solvers if the computational budget is small. We demonstrate that this results in significant formal verification speed-ups.

BSQ: Exploring Bit-Level Sparsity for Mixed-Precision Neural Network Quantization

- Huanrui Yang, Lin Duan, Yiran Chen, Hai Li
- abstract@open-review(Poster): Mixed-precision quantization can potentially achieve the optimal tradeoff between performance and compression rate of deep neural networks, and thus, have been widely investigated. However, it lacks a systematic method to determine the exact quantization scheme. Previous methods either examine only a small manually-designed search space or utilize a cumbersome neural architecture search to explore the vast search space. These approaches cannot lead to an optimal quantization scheme efficiently. This work proposes bit-level sparsity quantization (BSQ) to tackle the mixed-precision quantization from a new angle of inducing bit-level sparsity. We consider each bit of quantized weights as an independent trainable variable and introduce a differentiable bit-sparsity regularizer. BSQ can induce all-zero bits across a group of weight elements and realize the dynamic precision reduction, leading to a mixed-precision quantization scheme of the original model. Our method enables the exploration of the full mixed-precision space with a single gradient-based optimization process, with only one hyperparameter to tradeoff the performance and compression. BSQ achieves both higher accuracy and higher bit reduction on various model architectures on the CIFAR-10 and ImageNet datasets comparing to previous methods.

Dual-mode ASR: Unify and Improve Streaming ASR with Full-context Modeling

- Jiahui Yu, Wei Han, Anmol Gulati, Chung-Cheng Chiu, Bo Li, Tara N Sainath, Yonghui Wu, Ruoming Pang
- abstract@open-review(Poster): Streaming automatic speech recognition (ASR) aims to emit each hypothesized word as quickly and accurately as possible, while full-context ASR waits for the completion of a full speech utterance before emitting completed hypotheses. In this work, we propose a unified framework, Dual-mode ASR, to train a single end-to-end ASR model with shared weights for both streaming and full-context speech recognition. We show that the latency and accuracy of streaming ASR significantly benefit from weight sharing and joint training of full-context ASR, especially with inplace knowledge distillation during the training. The Dual-mode ASR framework can be applied to recent state-of-the-art convolution-based and transformer-based ASR networks. We present extensive experiments with two state-of-the-art ASR networks, ContextNet and Conformer, on two datasets, a widely used public dataset LibriSpeech and a large-scale dataset MultiDomain. Experiments and ablation studies demonstrate that Dual-mode ASR not only simplifies the workflow of training and deploying streaming and full-context ASR models, but also significantly improves both emission latency and recognition accuracy of streaming ASR. With Dual-mode ASR, we achieve new state-of-the-art streaming ASR results on both LibriSpeech and MultiDomain in terms of accuracy and latency.

Policy-Driven Attack: Learning to Query for Hard-label Black-box Adversarial Examples

- Ziang Yan, Yiwen Guo, Jian Liang, Changshui Zhang
- abstract@open-review(Poster): To craft black-box adversarial examples, adversaries need to query the victim model and take proper advantage of its feedback. Existing black-box attacks generally suffer from high query complexity, especially when only the top-1 decision (i.e., the hard-label prediction) of the victim model is available. In this paper, we propose a novel hard-label black-box attack named Policy-Driven Attack, to reduce the query complexity. Our core idea is to learn promising search directions of the adversarial examples using a well-designed policy network in a novel reinforcement learning formulation, in which the queries become more sensible. Experimental results demonstrate that our method can significantly reduce the query complexity in comparison with existing state-of-the-art hard-label black-box attacks on various image classification benchmark datasets. Code and models for reproducing our results are available at https://github.com/ZiangYan/pda.pytorch

Sequential Density Ratio Estimation for Simultaneous Optimization of Speed and Accuracy

- Akinori F Ebihara, Taiki Miyagawa, Kazuyuki Sakurai, Hitoshi Imaoka
- abstract@open-review(Spotlight): Classifying sequential data as early and as accurately as possible is a challenging yet critical problem, especially when a sampling cost is high. One algorithm that achieves this goal is the sequential probability ratio test (SPRT), which is known as Bayes-optimal: it can keep the expected number of data samples as small as possible, given the desired error upper-bound. However, the original SPRT makes two critical assumptions that limit its application in real-world scenarios: (i) samples are independently and identically distributed, and (ii) the likelihood of the data being derived from each class can be calculated precisely. Here, we propose the SPRT-TANDEM, a deep neural network-based SPRT algorithm that overcomes the above two obstacles. The SPRT-TANDEM sequentially estimates the log-likelihood ratio of two alternative hypotheses by leveraging a novel Loss function for Log-Likelihood Ratio estimation (LLLR) while allowing correlations up to $N (\in \mathbb{N})$ preceding samples. In tests on one original and two public video databases, Nosaic MNIST, UCF101, and SiW, the SPRT-TANDEM achieves statistically significantly better classification accuracy than other baseline classifiers, with a smaller number of data samples. The code and Nosaic MNIST are publicly available at https://github.com/TaikiMiyagawa/SPRT-TANDEM.

Uncertainty Sets for Image Classifiers using Conformal Prediction

- Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, Jitendra Malik
- abstract@open-review(Spotlight): Convolutional image classifiers can achieve high predictive accuracy, but quantifying their uncertainty remains an unresolved challenge, hindering their deployment in consequential settings. Existing uncertainty quantification techniques, such as Platt scaling, attempt to calibrate the network's probability estimates, but they do not have formal guarantees. We present an algorithm that modifies any classifier to output a predictive set containing the true label with a user-specified probability, such as 90%. The algorithm is simple and fast like Platt scaling, but provides a formal finite-sample coverage guarantee for every model and dataset. Our method modifies an existing conformal prediction algorithm to give more stable predictive sets by regularizing the small scores of unlikely classes after Platt scaling. In experiments on both Imagenet and Imagenet-V2 with ResNet-152 and other classifiers, our scheme outperforms existing approaches, achieving coverage with sets that are often factors of 5 to 10 smaller than a stand-alone Platt scaling baseline.

Graph Convolution with Low-rank Learnable Local Filters

- Xiuyuan Cheng, Zichen Miao, Qiang Qiu
- abstract@open-review(Spotlight): Geometric variations like rotation, scaling, and viewpoint changes pose a significant challenge to visual understanding. One common solution is to directly model certain intrinsic structures, e.g., using landmarks. However, it then becomes non-trivial to build effective deep models, especially when the underlying non-Euclidean grid is irregular and coarse. Recent deep models using graph convolutions provide an appropriate framework to handle such non-Euclidean data, but many of them, particularly those based on global graph Laplacians, lack expressiveness to capture local features required for representation of signals lying on the non-Euclidean grid. The current paper introduces a new type of graph convolution with learnable low-rank local filters, which is provably more expressive than previous spectral graph convolution methods. The model also provides a unified framework for both spectral and spatial graph convolutions. To improve model robustness, regularization by local graph Laplacians is introduced. The

representation stability against input graph data perturbation is theoretically proved, making use of the graph filter locality and the local graph regularization. Experiments on spherical mesh data, real-world facial expression recognition/skeleton-based action recognition data, and data with simulated graph noise show the empirical advantage of the proposed model.

PSTNet: Point Spatio-Temporal Convolution on Point Cloud Sequences

- Hehe Fan, Xin Yu, Yuhang Ding, Yi Yang, Mohan Kankanhalli
- abstract@open-review(Poster): Point cloud sequences are irregular and unordered in the spatial dimension while exhibiting regularities and order in the temporal dimension. Therefore, existing grid based convolutions for conventional video processing cannot be directly applied to spatio-temporal modeling of raw point cloud sequences. In this paper, we propose a point spatio-temporal (PST) convolution to achieve informative representations of point cloud sequences. The proposed PST convolution first disentangles space and time in point cloud sequences. Then, a spatial convolution is employed to capture the local structure of points in the 3D space, and a temporal convolution is used to model the dynamics of the spatial regions along the time dimension. Furthermore, we incorporate the proposed PST convolution into a deep network, namely PSTNet, to extract features of point cloud sequences in a hierarchical manner. Extensive experiments on widely-used 3D action recognition and 4D semantic segmentation datasets demonstrate the effectiveness of PSTNet to model point cloud sequences.

Network Pruning That Matters: A Case Study on Retraining Variants

- Duong Hoang Le, Binh-Son Hua
- abstract@open-review(Poster): Network pruning is an effective method to reduce the computational expense of over-parameterized neural networks for deployment on low-resource systems. Recent state-of-the-art techniques for retraining pruned networks such as weight rewinding and learning rate rewinding have been shown to outperform the traditional fine-tuning technique in recovering the lost accuracy (Renda et al., 2020), but so far it is unclear what accounts for such performance. In this work, we conduct extensive experiments to verify and analyze the uncanny effectiveness of learning rate rewinding. We find that the reason behind the success of learning rate rewinding is the usage of a large learning rate. Similar phenomenon can be observed in other learning rate schedules that involve large learning rates, e.g., the 1-cycle learning rate schedule (Smith et al., 2019). By leveraging the right learning rate schedule in retraining, we demonstrate a counter-intuitive phenomenon in that randomly pruned networks could even achieve better performance than methodically pruned networks (fine-tuned with the conventional approach). Our results emphasize the cruciality of the learning rate schedule in pruned network retraining - a detail often overlooked by practitioners during the implementation of network pruning.

EEC: Learning to Encode and Regenerate Images for Continual Learning

- Ali Ayub, Alan Wagner
- abstract@open-review(Poster): The two main impediments to continual learning are catastrophic forgetting and memory limitations on the storage of data. To cope with these challenges, we propose a novel, cognitively-inspired approach which trains autoencoders with Neural Style Transfer to encode and store images. Reconstructed images from encoded episodes are replayed when training the classifier model on a new task to avoid catastrophic forgetting. The loss function for the reconstructed images is weighted to reduce its effect during classifier training to cope with image degradation. When the system runs out of memory the encoded episodes are converted into centroids and covariance matrices, which are used to generate pseudo-images during classifier training, keeping classifier performance stable with less memory. Our approach increases classification accuracy by 13-17% over state-of-the-art methods on benchmark datasets, while requiring 78% less storage space.

Mind the Pad -- CNNs Can Develop Blind Spots

- Bilal Alsallakh, Narine Kokhlikyan, Vivek Miglani, Jun Yuan, Orion Reblitz-Richardson
- abstract@open-review(Spotlight): We show how feature maps in convolutional networks are susceptible to spatial bias. Due to a combination of architectural choices, the activation at certain locations is systematically elevated or weakened. The major source of this bias is the padding mechanism. Depending on several aspects of convolution arithmetic, this mechanism can apply the padding unevenly, leading to asymmetries in the learned weights. We demonstrate how such bias can be detrimental to certain tasks such as small object detection: the activation is suppressed if the stimulus lies in the impacted area, leading to blind spots and misdetection. We explore alternative padding methods and propose solutions for analyzing and mitigating spatial bias.

DARTS-: Robustly Stepping out of Performance Collapse Without Indicators

- Xiangxiang Chu, Xiaoxing Wang, Bo Zhang, Shun Lu, Xiaolin Wei, Junchi Yan
- abstract@open-review(Poster): Despite the fast development of differentiable architecture search (DARTS), it suffers from a standing instability issue regarding searching performance, which extremely limits its application. Existing robustifying methods draw clues from the outcome instead of finding out the causing factor. Various indicators such as Hessian eigenvalues are proposed as a signal of performance collapse, and the searching should be stopped once an indicator reaches a preset threshold. However, these methods tend to easily reject good architectures if thresholds are inappropriately set, let alone the searching is intrinsically noisy. In this paper, we undertake a more subtle and direct approach to resolve the collapse. We first demonstrate that skip connections with a learnable architectural coefficient can easily recover from a disadvantageous state and become dominant. We conjecture that skip connections profit too much from this privilege, hence causing the collapse for the derived model. Therefore, we propose to factor out this benefit with an auxiliary skip connection, ensuring a fairer competition for all operations. Extensive experiments on various datasets verify that our approach can substantially improve the robustness of DARTS. Our code is available at https://github.com/Meituan-AutoML/DARTS-

Stabilized Medical Image Attacks

- Gege Qi, Lijun GONG, Yibing Song, Kai Ma, Yefeng Zheng
- abstract@open-review(Spotlight): Convolutional Neural Networks (CNNs) have advanced existing medical systems for automatic disease diagnosis. However, a threat to these systems arises that adversarial attacks make CNNs vulnerable. Inaccurate diagnosis results make a negative influence on human healthcare. There is a need to investigate potential adversarial attacks to robustify deep medical diagnosis systems. On the other side, there are several modalities of medical images (e.g., CT, fundus, and endoscopic image) of which each type is significantly different from others. It is more challenging to generate adversarial perturbations for different types of medical images. In this paper, we propose an image-based medical adversarial attack method to consistently produce adversarial perturbations on medical images. The objective function of our method consists of a loss deviation term and a loss stabilization term. The loss deviation term increases the divergence between the CNN prediction of an adversarial example and its ground truth label. Meanwhile, the loss stabilization term ensures similar CNN predictions of this example and its smoothed input. From the perspective of the whole iterations for perturbation generation, the proposed loss stabilization term exhaustively searches the perturbation space to smooth the single spot for local optimum escape. We further analyze the KL-divergence of the proposed loss function and find that the loss stabilization term makes the perturbations updated towards a fixed objective spot while deviating from the ground truth. This stabilization ensures the proposed medical attack effective for different types of medical images while producing perturbations in small variance. Experiments on several medical image analysis benchmarks including the recent COVID-19 dataset show the stability of the proposed method.

BiPointNet: Binary Neural Network for Point Clouds

- Haotong Qin, Zhongang Cai, Mingyuan Zhang, Yifu Ding, Haiyu Zhao, Shuai Yi, Xianglong Liu, Hao Su
- abstract@open-review(Poster): To alleviate the resource constraint for real-time point cloud applications that run on edge devices, in this paper we present BiPointNet, the first model binarization approach for efficient deep learning on point clouds. We discover that the immense performance drop of binarized models for point clouds mainly stems from two challenges: aggregation-induced feature homogenization that leads to a degradation of information entropy, and scale distortion that hinders optimization and invalidates scale-sensitive structures. With theoretical justifications and in-depth analysis, our BiPointNet introduces Entropy-Maximizing Aggregation (EMA) to modulate the distribution before aggregation for the maximum information entropy, and Layer-wise Scale Recovery (LSR) to efficiently restore feature representation capacity. Extensive experiments show that BiPointNet outperforms existing binarization methods by convincing margins, at the level even comparable with the full precision counterpart. We highlight that our techniques are generic, guaranteeing significant improvements on various fundamental tasks and mainstream backbones. Moreover, BiPointNet gives an impressive 14.7× speedup and 18.9× storage saving on real-world resource-constrained devices.

AdaFuse: Adaptive Temporal Fusion Network for Efficient Action Recognition

- Yue Meng, Rameswar Panda, Chung-Ching Lin, Prasanna Sattigeri, Leonid Karlinsky, Kate Saenko, Aude Oliva, Rogerio Feris
- abstract@open-review(Poster): Temporal modelling is the key for efficient video action recognition. While understanding temporal information can improve recognition accuracy for dynamic actions, removing temporal redundancy and reusing past features can significantly save computation leading to efficient action recognition. In this paper, we introduce an adaptive temporal fusion network, called AdaFuse, that dynamically fuses channels from current and past feature maps for strong temporal modelling. Specifically, the necessary information from the historical convolution feature maps is fused with current pruned feature maps with the goal of improving both recognition accuracy and efficiency. In addition, we use a skipping operation to further reduce the computation cost of action recognition. Extensive experiments on SomethingV1 & V2, Jester and Mini-Kinetics show that our approach can achieve about 40% computation savings with comparable accuracy to state-of-the-art methods. The project page can be found at https://mengyuest.github.io/AdaFuse/

Generating Furry Cars: Disentangling Object Shape and Appearance across Multiple Domains

- Utkarsh Ojha, Krishna Kumar Singh, Yong Jae Lee
- abstract@open-review(Poster): We consider the novel task of learning disentangled representations of object shape and appearance across multiple domains (e.g., dogs and cars). The goal is to learn a generative model that learns an intermediate distribution, which borrows a subset of properties from each domain, enabling the generation of images that did not exist in any domain exclusively. This challenging problem requires an accurate disentanglement of object shape, appearance, and background from each domain, so that the appearance and shape factors from the two domains can be interchanged. We augment an existing approach that can disentangle factors within a single domain but struggles to do so across domains. Our key technical contribution is to represent object appearance with a differentiable histogram of visual features, and to optimize the generator so that two images with the same latent appearance factor but different latent shape factors produce similar histograms. On multiple multi-domain datasets, we demonstrate our method leads to accurate and consistent appearance and shape transfer across domains.

Pruning Neural Networks at Initialization: Why Are We Missing the Mark?

- Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, Michael Carbin
- abstract@open-review(Poster): Recent work has explored the possibility of pruning neural networks at initialization. We assess proposals for doing so: SNIP (Lee et al., 2019), GraSP (Wang et al., 2020), SynFlow (Tanaka et al., 2020), and magnitude pruning. Although these methods surpass the trivial baseline of random pruning, they remain below the accuracy of magnitude pruning after training, and we endeavor to understand why. We show that, unlike pruning after training, randomly shuffling the weights these methods prune within each layer or sampling new initial values preserves or improves accuracy. As such, the per-weight pruning decisions made by these methods can be replaced by a per-layer choice of the fraction of weights to prune. This property suggests broader challenges with the underlying pruning heuristics, the desire to prune at initialization, or both.

BRECQ: Pushing the Limit of Post-Training Quantization by Block Reconstruction

- Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, Shi Gu
- abstract@open-review(Poster): We study the challenging task of neural network quantization without end-to-end retraining, called Post-training Quantization (PTQ). PTQ usually requires a small subset of training data but produces less powerful quantized models than Quantization-Aware Training (QAT). In this work, we propose a novel PTQ framework, dubbed BRECQ, which pushes the limits of bitwidth in PTQ down to INT2 for the first time. BRECQ leverages the basic building blocks in neural networks and reconstructs them one-by-one. In a comprehensive theoretical study of the second-order error, we show that BRECQ achieves a good balance between cross-layer dependency and generalization error. To further employ the power of quantization, the mixed precision technique is incorporated in our framework by approximating the inter-layer and intra-layer sensitivity. Extensive experiments on various handcrafted and searched neural architectures are conducted for both image classification and object detection tasks. And for the first time we prove that, without bells and whistles, PTQ can attain 4-bit ResNet and MobileNetV2 comparable with QAT and enjoy 240 times faster production of quantized models. Codes are available at https://github.com/yhhhli/BRECQ.