Pay Attention to Features, Transfer Learn Faster CNNs

- Kafeng Wang, Xitong Gao, Yiren Zhao, Xingjian Li, Dejing Dou, Cheng-Zhong Xu
- abstract@open-review(Poster): Deep convolutional neural networks are now widely deployed in vision applications, but a limited size of training data can restrict their task performance. Transfer learning offers the chance for CNNs to learn with limited data samples by transferring knowledge from models pretrained on large datasets. Blindly transferring all learned features from the source dataset, however, brings unnecessary computation to CNNs on the target task. In this paper, we propose attentive feature distillation and selection (AFDS), which not only adjusts the strength of transfer learning regularization but also dynamically determines the important features to transfer. By deploying AFDS on ResNet-101, we achieved a state-of-the-art computation reduction at the same accuracy budget, outperforming all existing transfer learning methods. With a 10x MACs reduction budget, a ResNet-101 equipped with AFDS transfer learned from ImageNet to Stanford Dogs 120, can achieve an accuracy 11.07% higher than its best competitor.

Geom-GCN: Geometric Graph Convolutional Networks

- Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, Bo Yang
- abstract@open-review(Spotlight): Message-passing neural networks (MPNNs) have been successfully applied in a wide variety of applications in the real world. However, two fundamental weaknesses of MPNNs' aggregators limit their ability to represent graph-structured data: losing the structural information of nodes in neighborhoods and lacking the ability to capture long-range dependencies in disassortative graphs. Few studies have noticed the weaknesses from different perspectives. From the observations on classical neural network and network geometry, we propose a novel geometric aggregation scheme for graph neural networks to overcome the two weaknesses. The behind basic idea is the aggregation on a graph can benefit from a continuous space underlying the graph. The proposed aggregation scheme is permutation-invariant and consists of three modules, node embedding, structural neighborhood, and bi-level aggregation. We also present an implementation of the scheme in graph convolutional networks, termed Geom-GCN, to perform transductive learning on graphs. Experimental results show the proposed Geom-GCN achieved state-of-the-art performance on a wide range of open datasets of graphs.

Gradients as Features for Deep Representation Learning

- Fangzhou Mu, Yingyu Liang, Yin Li
- abstract@open-review(Poster): We address the challenging problem of deep representation learning -- the efficient adaption of a pre-trained deep network to different tasks. Specifically, we propose to explore gradient-based features. These features are gradients of the model parameters with respect to a task-specific loss given an input sample. Our key innovation is the design of a linear model that incorporates both gradient and activation of the pre-trained network. We demonstrate that our model provides a local linear approximation to an underlying deep model, and discuss important theoretical insights. Moreover, we present an efficient algorithm for the training and inference of our model without computing the actual gradients. Our method is evaluated across a number of representation-learning tasks on several datasets and using different network architectures. Strong results are obtained in all settings, and are well-aligned with our theoretical insights.

Monotonic Multihead Attention

- Xutai Ma, Juan Miguel Pino, James Cross, Liezl Puzon, Jiatao Gu
- abstract@open-review(Poster): Simultaneous machine translation models start generating a target sequence before they have encoded or read the source sequence. Recent approach for this task either apply a fixed policy on transformer, or a learnable monotonic attention on a weaker recurrent neural network based structure. In this paper, we propose a new attention mechanism, Monotonic Multihead Attention (MMA), which introduced the monotonic attention mechanism to multihead attention. We also introduced two novel interpretable approaches for latency control that are specifically designed for multiple attentions. We apply MMA to the simultaneous machine translation task and demonstrate better latency-quality tradeoffs compared to MILk, the previous state-of-the-art approach.

Massively Multilingual Sparse Word Representations

- Gábor Berend
- abstract@open-review(Poster): In this paper, we introduce Mamus for constructing multilingual sparse word representations. Our algorithm operates by determining a shared set of semantic units which get reutilized across languages, providing it a competitive edge both in terms of speed and evaluation performance. We demonstrate that our proposed algorithm behaves competitively to strong baselines through a series of rigorous experiments performed towards downstream applications spanning over dependency parsing, document classification and natural language inference. Additionally, our experiments relying on the QVEC-CCA evaluation score suggests that the proposed sparse word representations convey an increased interpretability as opposed to alternative approaches. Finally, we are releasing our multilingual sparse word representations for the 27 typologically diverse set of languages that we conducted our various experiments on.

Query-efficient Meta Attack to Deep Neural Networks

- Jiawei Du, Hu Zhang, Joey Tianyi Zhou, Yi Yang, Jiashi Feng
- abstract@open-review(Poster): Black-box attack methods aim to infer suitable attack patterns to targeted DNN models by only using output feedback of the models and the corresponding input queries. However, due to lack of prior and inefficiency in leveraging the query and feedback information, existing methods are mostly query-intensive for obtaining effective attack patterns. In this work, we propose a meta attack approach that is capable of attacking a targeted model with much fewer queries. Its high query-efficiency stems from effective utilization of meta learning approaches in learning generalizable prior abstraction from the previously observed attack patterns and exploiting such prior to help infer attack patterns from only a few queries and outputs. Extensive experiments on MNIST, CIFAR10 and tiny-Imagenet demonstrate that our meta-attack method can remarkably reduce the number of model queries without sacrificing the attack performance. Besides, the obtained meta attacker is not restricted to a particular model but can be used easily with a fast adaptive ability to attack a variety of models. Our code will be released to the public.

BREAKING CERTIFIED DEFENSES: SEMANTIC ADVERSARIAL EXAMPLES WITH SPOOFED ROBUSTNESS CERTIFICATES

- Amin Ghiasi, Ali Shafahi, Tom Goldstein
- abstract@open-review(Poster): Defenses against adversarial attacks can be classified into certified and non-certified. Certifiable defenses make networks robust within a certain ℓ_p-bounded radius, so that it is impossible for the adversary to make adversarial examples in the certificate bound. We present an attack that maintains the imperceptibility property of adversarial examples while being outside of the certified radius. Furthermore, the proposed "Shadow Attack" can fool certifiably robust networks by producing an imperceptible adversarial example that gets misclassified and produces a strong ``spoofed'' certificate.

An Exponential Learning Rate Schedule for Deep Learning

- Zhiyuan Li, Sanjeev Arora

- abstract@open-review(Spotlight): Intriguing empirical evidence exists that deep learning can work well with exotic schedules for varying the learning rate. This paper suggests that the phenomenon may be due to Batch Normalization or BN(Ioffe & Szegedy, 2015), which is ubiq- uitous and provides benefits in optimization and generalization across all standard architectures. The following new results are shown about BN with weight decay and momentum (in other words, the typical use case which was not considered in earlier theoretical analyses of stand-alone BN (Ioffe & Szegedy, 2015; Santurkar et al., 2018; Arora et al., 2018) • Training can be done using SGD with momentum and an exponentially in- creasing learning rate schedule, i.e., learning rate increases by some $(1 + \alpha)$ factor in every epoch for some $\alpha > 0$. (Precise statement in the paper.) To the best of our knowledge this is the first time such a rate schedule has been successfully used, let alone for highly successful architectures. As ex- pected, such training rapidly blows up network weights, but the net stays well-behaved due to normalization. • Mathematical explanation of the success of the above rate schedule: a rigor- ous proof that it is equivalent to the standard setting of BN + SGD + Standard Rate Tuning + Weight Decay + Momentum. This equivalence holds for other normalization layers as well, Group Normalization(Wu & He, 2018), Layer Normalization(Ba et al., 2016), Instance Norm(Ulyanov et al., 2016), etc. • A worked-out toy example illustrating the above linkage of hyper- parameters. Using either weight decay or BN alone reaches global minimum, but convergence fails when both are used.

Enabling Deep Spiking Neural Networks with Hybrid Conversion and Spike Timing Dependent Backpropagation

- Nitin Rathi, Gopalakrishnan Srinivasan, Priyadarshini Panda, Kaushik Roy
- abstract@open-review(Poster): Spiking Neural Networks (SNNs) operate with asynchronous discrete events (or spikes) which can potentially lead to higher energy-efficiency in neuromorphic hardware implementations. Many works have shown that an SNN for inference can be formed by copying the weights from a trained Artificial Neural Network (ANN) and setting the firing threshold for each layer as the maximum input received in that layer. These type of converted SNNs require a large number of time steps to achieve competitive accuracy which diminishes the energy savings. The number of time steps can be reduced by training SNNs with spike-based backpropagation from scratch, but that is computationally expensive and slow. To address these challenges, we present a computationally-efficient training technique for deep SNNs. We propose a hybrid training methodology: 1) take a converted SNN and use its weights and thresholds as an initialization step for spike-based backpropagation, and 2) perform incremental spike-timing dependent backpropagation (STDB) on this carefully initialized network to obtain an SNN that converges within few epochs and requires fewer time steps for input processing. STDB is performed with a novel surrogate gradient function defined using neuron's spike time. The weight update is proportional to the difference in spike timing between the current time step and the most recent time step the neuron generated an output spike. The SNNs trained with our hybrid conversion-and-STDB training perform at $10{\times}{-}25{\times}$ fewer number of time steps and achieve similar accuracy compared to purely converted SNNs. The proposed training methodology converges in less than 20 epochs of spike-based backpropagation for most standard image classification datasets, thereby greatly reducing the training complexity compared to training SNNs from scratch. We perform experiments on CIFAR-10, CIFAR-100 and ImageNet datasets for both VGG and ResNet architectures. We achieve top-1 accuracy of 65.19% for ImageNet dataset on SNN with 250 time steps, which is $10{\times}$ faster compared to converted SNNs with similar accuracy.

How to 0wn the NAS in Your Spare Time

- Sanghyun Hong, Michael Davinroy, Yiğitcan Kaya, Dana Dachman-Soled, Tudor Dumitraş
- abstract@open-review(Poster): New data processing pipelines and novel network architectures increasingly drive the success of deep learning. In consequence, the industry considers top-performing architectures as intellectual property and devotes considerable computational resources to discovering such architectures through neural architecture search (NAS). This provides an incentive for adversaries to steal these novel architectures; when used in the cloud, to provide Machine Learning as a Service (MLaaS), the adversaries also have an opportunity to reconstruct the architectures by exploiting a range of hardware side-channels. However, it is challenging to reconstruct novel architectures and pipelines without knowing the computational graph (e.g., the layers, branches or skip connections), the architectural parameters (e.g., the number of filters in a convolutional layer) or the specific pre-processing steps (e.g. embeddings). In this paper, we design an algorithm that reconstructs the key components of a novel deep learning system by exploiting a small amount of information leakage from a cache side-channel attack, Flush+Reload. We use Flush+Reload to infer the trace of computations and the timing for each computation. Our algorithm then generates candidate computational graphs from the trace and eliminates incompatible candidates through a parameter estimation process. We implement our algorithm in PyTorch and Tensorflow. We demonstrate experimentally that we can reconstruct MalConv, a novel data pre-processing pipeline for malware detection, and ProxylessNAS-CPU, a novel network architecture for the ImageNet classification optimized to run on CPUs, without knowing the architecture family. In both cases, we achieve 0% error. These results suggest hardware side channels are a practical attack vector against MLaaS, and more efforts should be devoted to understanding their impact on the security of deep learning systems.

The Shape of Data: Intrinsic Distance for Data Distributions

- Anton Tsitsulin, Marina Munkhoeva, Davide Mottin, Panagiotis Karras, Alex Bronstein, Ivan Oseledets, Emmanuel Mueller
- abstract@open-review(Poster): The ability to represent and compare machine learning models is crucial in order to quantify subtle model changes, evaluate generative models, and gather insights on neural network architectures. Existing techniques for comparing data distributions focus on global data properties such as mean and covariance; in that sense, they are extrinsic and uni-scale. We develop a first-of-its-kind intrinsic and multi-scale method for characterizing and comparing data manifolds, using a lower-bound of the spectral variant of the Gromov-Wasserstein inter-manifold distance, which compares all data moments. In a thorough experimental study, we demonstrate that our method effectively discerns the structure of data manifolds even on unaligned data of different dimensionalities; moreover, we showcase its efficacy in evaluating the quality of generative models.

Understanding Generalization in Recurrent Neural Networks

- Zhuozhuo Tu, Fengxiang He, Dacheng Tao
- abstract@open-review(Poster): In this work, we develop the theory for analyzing the generalization performance of recurrent neural networks. We first present a new generalization bound for recurrent neural networks based on matrix 1-norm and Fisher-Rao norm. The definition of Fisher-Rao norm relies on a structural lemma about the gradient of RNNs. This new generalization bound assumes that the covariance matrix of the input data is positive definite, which might limit its use in practice. To address this issue, we propose to add random noise to the input data and prove a generalization bound for training with random noise, which is an extension of the former one. Compared with existing results, our generalization bounds have no explicit dependency on the size of networks. We also discover that Fisher-Rao norm for RNNs can be interpreted as a measure of gradient, and incorporating this gradient measure not only can tighten the bound, but allows us to build a relationship between generalization and trainability. Based on the bound, we theoretically analyze the effect of covariance of features on generalization of RNNs and discuss how weight decay and gradient clipping in the training can help improve generalization.

Conservative Uncertainty Estimation By Fitting Prior Networks

- Kamil Ciosek, Vincent Fortuin, Ryota Tomioka, Katja Hofmann, Richard Turner
- abstract@open-review(Poster): Obtaining high-quality uncertainty estimates is essential for many applications of deep neural networks. In this paper, we theoretically justify a scheme for estimating uncertainties, based on sampling from a prior distribution. Crucially, the uncertainty estimates are shown to be conservative in the sense that they never underestimate a posterior uncertainty obtained by a hypothetical Bayesian algorithm. We also show concentration, implying that the uncertainty estimates converge to zero as we get more data. Uncertainty estimates obtained from random priors can be adapted to any deep network architecture and trained using standard supervised learning pipelines. We provide experimental evaluation of random priors on calibration and out-of-distribution detection on typical computer vision tasks, demonstrating that they outperform deep ensembles in practice.

NAS-Bench-1Shot1: Benchmarking and Dissecting One-shot Neural Architecture Search

- Arber Zela, Julien Siems, Frank Hutter
- abstract@open-review(Poster): One-shot neural architecture search (NAS) has played a crucial role in making NAS methods computationally feasible in practice. Nevertheless, there is still a lack of understanding on how these weight-sharing algorithms exactly work due to the many factors controlling the dynamics of the process. In order to allow a scientific study of these components, we introduce a general framework for one-shot NAS that can be instantiated to many recently-introduced variants and introduce a general benchmarking framework that draws on the recent large-scale tabular benchmark NAS-Bench-101 for cheap anytime evaluations of one-shot NAS methods. To showcase the framework, we compare several state-of-the-art one-shot NAS methods, examine how sensitive they are to their hyperparameters and how they can be improved by tuning their hyperparameters, and compare their performance to that of blackbox optimizers for NAS-Bench-101.

Learning to Coordinate Manipulation Skills via Skill Behavior Diversification

- Youngwoon Lee, Jingyun Yang, Joseph J. Lim
- abstract@open-review(Poster): When mastering a complex manipulation task, humans often decompose the task into sub-skills of their body parts, practice the sub-skills independently, and then execute the sub-skills together. Similarly, a robot with multiple end-effectors can perform complex tasks by coordinating sub-skills of each end-effector. To realize temporal and behavioral coordination of skills, we propose a modular framework that first individually trains sub-skills of each end-effector with skill behavior diversification, and then learns to coordinate end-effectors using diverse behaviors of the skills. We demonstrate that our proposed framework is able to efficiently coordinate skills to solve challenging collaborative control tasks such as picking up a long bar, placing a block inside a container while pushing the container with two robot arms, and pushing a box with two ant agents. Videos and code are available at https://clvrai.com/coordination

Robust Subspace Recovery Layer for Unsupervised Anomaly Detection

- Chieh-Hsin Lai, Dongmian Zou, Gilad Lerman
- abstract@open-review(Poster): We propose a neural network for unsupervised anomaly detection with a novel robust subspace recovery layer (RSR layer). This layer seeks to extract the underlying subspace from a latent representation of the given data and removes outliers that lie away from this subspace. It is used within an autoencoder. The encoder maps the data into a latent space, from which the RSR layer extracts the subspace. The decoder then smoothly maps back the underlying subspace to a ``manifold" close to the original inliers. Inliers and outliers are distinguished according to the distances between the original and mapped positions (small for inliers and large for outliers). Extensive numerical experiments with both image and document datasets demonstrate state-of-the-art precision and recall.

Learning Nearly Decomposable Value Functions Via Communication Minimization

- Tonghan Wang, *Jianhao Wang*, Chongyi Zheng, Chongjie Zhang
- abstract@open-review(Poster): Reinforcement learning encounters major challenges in multi-agent settings, such as scalability and non-stationarity. Recently, value function factorization learning emerges as a promising way to address these challenges in collaborative multi-agent systems. However, existing methods have been focusing on learning fully decentralized value functions, which are not efficient for tasks requiring communication. To address this limitation, this paper presents a novel framework for learning nearly decomposable Q-functions (NDQ) via communication minimization, with which agents act on their own most of the time but occasionally send messages to other agents in order for effective coordination. This framework hybridizes value function factorization learning and communication learning by introducing two information-theoretic regularizers. These regularizers are maximizing mutual information between agents' action selection and communication messages while minimizing the entropy of messages between agents. We show how to optimize these regularizers in a way that is easily integrated with existing value function factorization methods such as QMIX. Finally, we demonstrate that, on the StarCraft unit micromanagement benchmark, our framework significantly outperforms baseline methods and allows us to cut off more than 80% of communication without sacrificing the performance. The videos of our experiments are available at https://sites.google.com/view/ndq.

Extreme Classification via Adversarial Softmax Approximation

- Robert Bamler, Stephan Mandt
- abstract@open-review(Poster): Training a classifier over a large number of classes, known as 'extreme classification', has become a topic of major interest with applications in technology, science, and e-commerce. Traditional softmax regression induces a gradient cost proportional to the number of classes C, which often is prohibitively expensive. A popular scalable softmax approximation relies on uniform negative sampling, which suffers from slow convergence due a poor signal-to-noise ratio. In this paper, we propose a simple training method for drastically enhancing the gradient signal by drawing negative samples from an adversarial model that mimics the data distribution. Our contributions are three-fold: (i) an adversarial sampling mechanism that produces negative samples at a cost only logarithmic in C, thus still resulting in cheap gradient updates; (ii) a mathematical proof that this adversarial sampling minimizes the gradient variance while any bias due to non-uniform sampling can be removed; (iii) experimental results on large scale data sets that show a reduction of the training time by an order of magnitude relative to several competitive baselines.

Information Geometry of Orthogonal Initializations and Training

- Piotr Aleksander Sokół, Il Memming Park
- abstract@open-review(Poster): Recently mean field theory has been successfully used to analyze properties of wide, random neural networks. It gave rise to a prescriptive theory for initializing feed-forward neural networks with orthogonal weights, which ensures that both the forward propagated activations and the backpropagated gradients are near (\ell_2) isometries and as a consequence training is orders of magnitude faster. Despite strong empirical performance, the mechanisms by which critical initializations confer an advantage in the optimization of deep neural networks are poorly understood. Here we show a novel connection between the maximum curvature of the optimization landscape (gradient smoothness) as measured by the Fisher information matrix (FIM) and the spectral radius of the input-output Jacobian, which partially explains why more isometric networks can train much faster. Furthermore, given that orthogonal weights are necessary to ensure that gradient norms are approximately preserved at initialization, we experimentally investigate the benefits of maintaining orthogonality throughout training, and we conclude that manifold optimization of weights performs well regardless of the smoothness of the gradients. Moreover, we observe a surprising yet robust behavior of highly isometric initializations --- even though such networks have a lower FIM condition number \emph{at initialization}, and therefore by analogy to convex functions should be easier to optimize, experimentally they prove to be much harder to train with stochastic gradient descent. We conjecture the FIM condition number plays a non-trivial role in the optimization.

Mixed Precision DNNs: All you need is a good parametrization

- Stefan Uhlich, Lukas Mauch, Fabien Cardinaux, Kazuki Yoshiyama, Javier Alonso Garcia, Stephen Tiedemann, Thomas Kemp, Akira Nakamura
- abstract@open-review(Poster): Efficient deep neural network (DNN) inference on mobile or embedded devices typically involves quantization of the network parameters and activations. In particular, mixed precision networks achieve better performance than networks with homogeneous bitwidth for the same size constraint. Since choosing the optimal bitwidths is not straight forward, training methods, which can learn them, are desirable. Differentiable

quantization with straight-through gradients allows to learn the quantizer's parameters using gradient methods. We show that a suited parametrization of the quantizer is the key to achieve a stable training and a good final performance. Specifically, we propose to parametrize the quantizer with the step size and dynamic range. The bitwidth can then be inferred from them. Other parametrizations, which explicitly use the bitwidth, consistently perform worse. We confirm our findings with experiments on CIFAR-10 and ImageNet and we obtain mixed precision DNNs with learned quantization parameters, achieving state-of-the-art performance.

PROGRESSIVE LEARNING AND DISENTANGLEMENT OF HIERARCHICAL REPRESENTATIONS

- Zhiyuan Li, Jaideep Vitthal Murkute, Prashnna Kumar Gyawali, Linwei Wang
- abstract@open-review(Spotlight): Learning rich representation from data is an important task for deep generative models such as variational auto-encoder (VAE). However, by extracting high-level abstractions in the bottom-up inference process, the goal of preserving all factors of variations for top-down generation is compromised. Motivated by the concept of “starting small”, we present a strategy to progressively learn independent hierarchical representations from high- to low-levels of abstractions. The model starts with learning the most abstract representation, and then progressively grow the network architecture to introduce new representations at different levels of abstraction. We quantitatively demonstrate the ability of the presented model to improve disentanglement in comparison to existing works on two benchmark datasets using three disentanglement metrics, including a new metric we proposed to complement the previously-presented metric of mutual information gap. We further present both qualitative and quantitative evidence on how the progression of learning improves disentangling of hierarchical representations. By drawing on the respective advantage of hierarchical representation learning and progressive learning, this is to our knowledge the first attempt to improve disentanglement by progressively growing the capacity of VAE to learn hierarchical representations.

Co-Attentive Equivariant Neural Networks: Focusing Equivariance On Transformations Co-Occurring in Data

- David W. Romero, Mark Hoogendoorn
- abstract@open-review(Poster): Equivariance is a nice property to have as it produces much more parameter efficient neural architectures and preserves the structure of the input through the feature mapping. Even though some combinations of transformations might never appear (e.g. an upright face with a horizontal nose), current equivariant architectures consider the set of all possible transformations in a transformation group when learning feature representations. Contrarily, the human visual system is able to attend to the set of relevant transformations occurring in the environment and utilizes this information to assist and improve object recognition. Based on this observation, we modify conventional equivariant feature mappings such that they are able to attend to the set of co-occurring transformations in data and generalize this notion to act on groups consisting of multiple symmetries. We show that our proposed co-attentive equivariant neural networks consistently outperform conventional rotation equivariant and rotation & reflection equivariant neural networks on rotated MNIST and CIFAR-10.

Deep Orientation Uncertainty Learning based on a Bingham Loss

- Igor Gilitschenski, Roshni Sahoo, Wilko Schwarting, Alexander Amini, Sertac Karaman, Daniela Rus
- abstract@open-review(Poster): Reasoning about uncertain orientations is one of the core problems in many perception tasks such as object pose estimation or motion estimation. In these scenarios, poor illumination conditions, sensor limitations, or appearance invariance may result in highly uncertain estimates. In this work, we propose a novel learning-based representation for orientation uncertainty. By characterizing uncertainty over unit quaternions with the Bingham distribution, we formulate a loss that naturally captures the antipodal symmetry of the representation. We discuss the interpretability of the learned distribution parameters and demonstrate the feasibility of our approach on several challenging real-world pose estimation tasks involving uncertain orientations.

Reconstructing continuous distributions of 3D protein structure from cryo-EM images

- Ellen D. Zhong, Tristan Bepler, Joseph H. Davis, Bonnie Berger
- abstract@open-review(Spotlight): Cryo-electron microscopy (cryo-EM) is a powerful technique for determining the structure of proteins and other macromolecular complexes at near-atomic resolution. In single particle cryo-EM, the central problem is to reconstruct the 3D structure of a macromolecule from 10^{4-7} noisy and randomly oriented 2D projection images. However, the imaged protein complexes may exhibit structural variability, which complicates reconstruction and is typically addressed using discrete clustering approaches that fail to capture the full range of protein dynamics. Here, we introduce a novel method for cryo-EM reconstruction that extends naturally to modeling continuous generative factors of structural heterogeneity. This method encodes structures in Fourier space using coordinate-based deep neural networks, and trains these networks from unlabeled 2D cryo-EM images by combining exact inference over image orientation with variational inference for structural heterogeneity. We demonstrate that the proposed method, termed cryoDRGN, can perform ab-initio reconstruction of 3D protein complexes from simulated and real 2D cryo-EM image data. To our knowledge, cryoDRGN is the first neural network-based approach for cryo-EM reconstruction and the first end-to-end method for directly reconstructing continuous ensembles of protein structures from cryo-EM images.

Generalization of Two-layer Neural Networks: An Asymptotic Viewpoint

- Jimmy Ba, Murat Erdogdu, Taiji Suzuki, Denny Wu, Tianzong Zhang
- abstract@open-review(Spotlight): This paper investigates the generalization properties of two-layer neural networks in high-dimensions, i.e. when the number of samples n, features d, and neurons h tend to infinity at the same rate. Specifically, we derive the exact population risk of the unregularized least squares regression problem with two-layer neural networks when either the first or the second layer is trained using a gradient flow under different initialization setups. When only the second layer coefficients are optimized, we recover the \textit{double descent} phenomenon: a cusp in the population risk appears at $h\approx n$ and further overparameterization decreases the risk. In contrast, when the first layer weights are optimized, we highlight how different scales of initialization lead to different inductive bias, and show that the resulting risk is \textit{independent} of overparameterization. Our theoretical and experimental results suggest that previously studied model setups that provably give rise to \textit{double descent} might not translate to optimizing two-layer neural networks.

Critical initialisation in continuous approximations of binary neural networks

- George Stamatescu, Federica Gerace, Carlo Lucibello, Ian Fuss, Langford White
- abstract@open-review(Poster): The training of stochastic neural network models with binary (±1) weights and activations via continuous surrogate networks is investigated. We derive new surrogates using a novel derivation based on writing the stochastic neural network as a Markov chain. This derivation also encompasses existing variants of the surrogates presented in the literature. Following this, we theoretically study the surrogates at initialisation. We derive, using mean field theory, a set of scalar equations describing how input signals propagate through the randomly initialised networks. The equations reveal whether so-called critical initialisations exist for each surrogate network, where the network can be trained to arbitrary depth. Moreover, we predict theoretically and confirm numerically, that common weight initialisation schemes used in standard continuous networks, when applied to the mean values of the stochastic binary weights, yield poor training performance. This study shows that, contrary to common intuition, the means of the stochastic binary weights should be initialised close to ± 1, for deeper networks to be trainable.

Sub-policy Adaptation for Hierarchical Reinforcement Learning

- Alexander Li, Carlos Florensa, Ignasi Clavera, Pieter Abbeel
- abstract@open-review(Poster): Hierarchical reinforcement learning is a promising approach to tackle long-horizon decision-making problems with sparse rewards. Unfortunately, most methods still decouple the lower-level skill acquisition process and the training of a higher level that controls the skills in a new task. Leaving the skills fixed can lead to significant sub-optimality in the transfer setting. In this work, we propose a novel algorithm to discover a set of skills, and continuously adapt them along with the higher level even when training on a new task. Our main contributions are two-fold. First, we derive a new hierarchical policy gradient with an unbiased latent-dependent baseline, and we introduce Hierarchical Proximal Policy Optimization (HiPPO), an on-policy method to efficiently train all levels of the hierarchy jointly. Second, we propose a method of training time-abstractions that improves the robustness of the obtained skills to environment changes. Code and videos are available at sites.google.com/view/hippo-rl.

Episodic Reinforcement Learning with Associative Memory

- Guangxiang Zhu, *Zichuan Lin*, Guangwen Yang, Chongjie Zhang
- abstract@open-review(Poster): Sample efficiency has been one of the major challenges for deep reinforcement learning. Non-parametric episodic control has been proposed to speed up parametric reinforcement learning by rapidly latching on previously successful policies. However, previous work on episodic reinforcement learning neglects the relationship between states and only stored the experiences as unrelated items. To improve sample efficiency of reinforcement learning, we propose a novel framework, called Episodic Reinforcement Learning with Associative Memory (ERLAM), which associates related experience trajectories to enable reasoning effective strategies. We build a graph on top of states in memory based on state transitions and develop a reverse-trajectory propagation strategy to allow rapid value propagation through the graph. We use the non-parametric associative memory as early guidance for a parametric reinforcement learning model. Results on navigation domain and Atari games show our framework achieves significantly higher sample efficiency than state-of-the-art episodic reinforcement learning models.

Depth-Width Trade-offs for ReLU Networks via Sharkovsky's Theorem

- Vaggos Chatziafratis, Sai Ganesh Nagarajan, Ioannis Panageas, Xiao Wang
- abstract@open-review(Spotlight): Understanding the representational power of Deep Neural Networks (DNNs) and how their structural properties (e.g., depth, width, type of activation unit) affect the functions they can compute, has been an important yet challenging question in deep learning and approximation theory. In a seminal paper, Telgarsky high- lighted the benefits of depth by presenting a family of functions (based on sim- ple triangular waves) for which DNNs achieve zero classification error, whereas shallow networks with fewer than exponentially many nodes incur constant error. Even though Telgarsky's work reveals the limitations of shallow neural networks, it doesn't inform us on why these functions are difficult to represent and in fact he states it as a tantalizing open question to characterize those functions that cannot be well-approximated by smaller depths. In this work, we point to a new connection between DNNs expressivity and Sharkovsky's Theorem from dynamical systems, that enables us to characterize the depth-width trade-offs of ReLU networks for representing functions based on the presence of a generalized notion of fixed points, called periodic points (a fixed point is a point of period 1). Motivated by our observation that the triangle waves used in Telgarsky's work contain points of period 3 – a period that is special in that it implies chaotic behaviour based on the celebrated result by Li-Yorke – we proceed to give general lower bounds for the width needed to represent periodic functions as a function of the depth. Technically, the crux of our approach is based on an eigenvalue analysis of the dynamical systems associated with such functions.

Energy-based models for atomic-resolution protein conformations

- Yilun Du, Joshua Meier, Jerry Ma, Rob Fergus, Alexander Rives
- abstract@open-review(Spotlight): We propose an energy-based model (EBM) of protein conformations that operates at atomic scale. The model is trained solely on crystallized protein data. By contrast, existing approaches for scoring conformations use energy functions that incorporate knowledge of physical principles and features that are the complex product of several decades of research and tuning. To evaluate the model, we benchmark on the rotamer recovery task, the problem of predicting the conformation of a side chain from its context within a protein structure, which has been used to evaluate energy functions for protein design. The model achieves performance close to that of the Rosetta energy function, a state-of-the-art method widely used in protein structure prediction and design. An investigation of the model's outputs and hidden representations finds that it captures physicochemical properties relevant to protein energy.

Federated Learning with Matched Averaging

- Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, Yasaman Khazaeni
- abstract@open-review(Talk): Federated learning allows edge devices to collaboratively learn a shared model while keeping the training data on device, decoupling the ability to do model training from the need to store the data in the cloud. We propose Federated matched averaging (FedMA) algorithm designed for federated learning of modern neural network architectures e.g. convolutional neural networks (CNNs) and LSTMs. FedMA constructs the shared global model in a layer-wise manner by matching and averaging hidden elements (i.e. channels for convolution layers; hidden states for LSTM; neurons for fully connected layers) with similar feature extraction signatures. Our experiments indicate that FedMA not only outperforms popular state-of-the-art federated learning algorithms on deep CNN and LSTM architectures trained on real world datasets, but also reduces the overall communication burden.

Pitfalls of In-Domain Uncertainty Estimation and Ensembling in Deep Learning

- Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, Dmitry Vetrov
- abstract@open-review(Poster): Uncertainty estimation and ensembling methods go hand-in-hand. Uncertainty estimation is one of the main benchmarks for assessment of ensembling performance. At the same time, deep learning ensembles have provided state-of-the-art results in uncertainty estimation. In this work, we focus on in-domain uncertainty for image classification. We explore the standards for its quantification and point out pitfalls of existing metrics. Avoiding these pitfalls, we perform a broad study of different ensembling techniques. To provide more insight in this study, we introduce the deep ensemble equivalent score (DEE) and show that many sophisticated ensembling techniques are equivalent to an ensemble of only few independently trained networks in terms of test performance.

DiffTaichi: Differentiable Programming for Physical Simulation

- Yuanming Hu, Luke Anderson, Tzu-Mao Li, Qi Sun, Nathan Carr, Jonathan Ragan-Kelley, Fredo Durand
- abstract@open-review(Poster): We present DiffTaichi, a new differentiable programming language tailored for building high-performance differentiable physical simulators. Based on an imperative programming language, DiffTaichi generates gradients of simulation steps using source code transformations that preserve arithmetic intensity and parallelism. A light-weight tape is used to record the whole simulation program structure and replay the gradient kernels in a reversed order, for end-to-end backpropagation. We demonstrate the performance and productivity of our language in gradient-based learning and optimization tasks on 10 different physical simulators. For example, a differentiable elastic object simulator written in our language is 4.2x shorter than the hand-engineered CUDA version yet runs as fast, and is 188x faster than the TensorFlow implementation. Using our differentiable programs, neural network controllers are typically optimized within only tens of iterations.

A Mutual Information Maximization Perspective of Language Representation Learning

- Lingpeng Kong, Cyprien de Masson d'Autume, Lei Yu, Wang Ling, Zihang Dai, Dani Yogatama
- abstract@open-review(Spotlight): We show state-of-the-art word representation learning methods maximize an objective function that is a lower bound on the mutual information between different parts of a word sequence (i.e., a sentence). Our formulation provides an alternative perspective that unifies classical word embedding models (e.g., Skip-gram) and modern contextual embeddings (e.g., BERT, XLNet). In addition to enhancing our theoretical understanding of these methods, our derivation leads to a principled framework that can be used to construct new self-supervised tasks. We provide an example by drawing inspirations from related methods based on mutual information maximization that have been successful in computer vision, and introduce a simple self-supervised objective that maximizes the mutual information between a global sentence representation and n-grams in the sentence. Our analysis offers a holistic view of representation learning methods to transfer knowledge and translate progress across multiple domains (e.g., natural language processing, computer vision, audio processing).

Domain Adaptive Multibranch Networks

- Róger Bermúdez-Chacón, Mathieu Salzmann, Pascal Fua
- abstract@open-review(Poster): We tackle unsupervised domain adaptation by accounting for the fact that different domains may need to be processed differently to arrive to a common feature representation effective for recognition. To this end, we introduce a deep learning framework where each domain undergoes a different sequence of operations, allowing some, possibly more complex, domains to go through more computations than others. This contrasts with state-of-the-art domain adaptation techniques that force all domains to be processed with the same series of operations, even when using multi-stream architectures whose parameters are not shared. As evidenced by our experiments, the greater flexibility of our method translates to higher accuracy. Furthermore, it allows us to handle any number of domains simultaneously.

Unbiased Contrastive Divergence Algorithm for Training Energy-Based Latent Variable Models

- Yixuan Qiu, Lingsong Zhang, Xiao Wang
- abstract@open-review(Spotlight): The contrastive divergence algorithm is a popular approach to training energy-based latent variable models, which has been widely used in many machine learning models such as the restricted Boltzmann machines and deep belief nets. Despite its empirical success, the contrastive divergence algorithm is also known to have biases that severely affect its convergence. In this article we propose an unbiased version of the contrastive divergence algorithm that completely removes its bias in stochastic gradient methods, based on recent advances on unbiased Markov chain Monte Carlo methods. Rigorous theoretical analysis is developed to justify the proposed algorithm, and numerical experiments show that it significantly improves the existing method. Our findings suggest that the unbiased contrastive divergence algorithm is a promising approach to training general energy-based latent variable models.

Differentiable Reasoning over a Virtual Knowledge Base

- Bhuwan Dhingra, Manzil Zaheer, Vidhisha Balachandran, Graham Neubig, Ruslan Salakhutdinov, William W. Cohen
- abstract@open-review(Talk): We consider the task of answering complex multi-hop questions using a corpus as a virtual knowledge base (KB). In particular, we describe a neural module, DrKIT, that traverses textual data like a KB, softly following paths of relations between mentions of entities in the corpus. At each step the module uses a combination of sparse-matrix TFIDF indices and a maximum inner product search (MIPS) on a special index of contextual representations of the mentions. This module is differentiable, so the full system can be trained end-to-end using gradient based methods, starting from natural language inputs. We also describe a pretraining scheme for the contextual representation encoder by generating hard negative examples using existing knowledge bases. We show that DrKIT improves accuracy by 9 points on 3-hop questions in the MetaQA dataset, cutting the gap between text-based and KB-based state-of-the-art by 70%. On HotpotQA, DrKIT leads to a 10% improvement over a BERT-based re-ranking approach to retrieving the relevant passages required to answer a question. DrKIT is also very efficient, processing up to 10-100x more queries per second than existing multi-hop systems.

Making Sense of Reinforcement Learning and Probabilistic Inference

- Brendan O'Donoghue, Ian Osband, Catalin Ionescu
- abstract@open-review(Spotlight): Reinforcement learning (RL) combines a control problem with statistical estimation: The system dynamics are not known to the agent, but can be learned through experience. A recent line of research casts ‘RL as inference’ and suggests a particular framework to generalize the RL problem as probabilistic inference. Our paper surfaces a key shortcoming in that approach, and clarifies the sense in which RL can be coherently cast as an inference problem. In particular, an RL agent must consider the effects of its actions upon future rewards and observations: The exploration-exploitation tradeoff. In all but the most simple settings, the resulting inference is computationally intractable so that practical RL algorithms must resort to approximation. We demonstrate that the popular ‘RL as inference’ approximation can perform poorly in even very basic problems. However, we show that with a small modification the framework does yield algorithms that can provably perform well, and we show that the resulting algorithm is equivalent to the recently proposed K-learning, which we further connect with Thompson sampling.

Prediction Poisoning: Towards Defenses Against DNN Model Stealing Attacks

- Tribhuvanesh Orekondy, Bernt Schiele, Mario Fritz
- abstract@open-review(Poster): High-performance Deep Neural Networks (DNNs) are increasingly deployed in many real-world applications e.g., cloud prediction APIs. Recent advances in model functionality stealing attacks via black-box access (i.e., inputs in, predictions out) threaten the business model of such applications, which require a lot of time, money, and effort to develop. Existing defenses take a passive role against stealing attacks, such as by truncating predicted information. We find such passive defenses ineffective against DNN stealing attacks. In this paper, we propose the first defense which actively perturbs predictions targeted at poisoning the training objective of the attacker. We find our defense effective across a wide range of challenging datasets and DNN model stealing attacks, and additionally outperforms existing defenses. Our defense is the first that can withstand highly accurate model stealing attacks for tens of thousands of queries, amplifying the attacker's error rate up to a factor of 85$\times$ with minimal impact on the utility for benign users.

SCALOR: Generative World Models with Scalable Object Representations

- Jindong Jiang, *Sepehr Janghorbani*, Gerard De Melo, Sungjin Ahn
- abstract@open-review(Poster): Scalability in terms of object density in a scene is a primary challenge in unsupervised sequential object-oriented representation learning. Most of the previous models have been shown to work only on scenes with a few objects. In this paper, we propose SCALOR, a probabilistic generative world model for learning SCALable Object-oriented Representation of a video. With the proposed spatially parallel attention and proposal-rejection mechanisms, SCALOR can deal with orders of magnitude larger numbers of objects compared to the previous state-of-the-art models. Additionally, we introduce a background module that allows SCALOR to model complex dynamic backgrounds as well as many foreground objects in the scene. We demonstrate that SCALOR can deal with crowded scenes containing up to a hundred objects while jointly modeling complex dynamic backgrounds. Importantly, SCALOR is the first unsupervised object representation model shown to work for natural scenes containing several tens of moving objects.

Neural tangent kernels, transportation mappings, and universal approximation

- Ziwei Ji, Matus Telgarsky, Ruicheng Xian
- abstract@open-review(Poster): This paper establishes rates of universal approximation for the shallow neural tangent kernel (NTK): network weights are only allowed microscopic changes from random initialization, which entails that activations are mostly unchanged, and the network is nearly equivalent to its linearization. Concretely, the paper has two main contributions: a generic scheme to approximate functions with the NTK by sampling from transport mappings between the initial weights and their desired values, and the construction of transport mappings via Fourier transforms. Regarding the first contribution, the proof scheme provides another perspective on how the NTK regime arises from rescaling: redundancy in the weights due to resampling allows individual weights to be scaled down. Regarding the second contribution, the most notable transport mapping asserts that roughly $1 / \delta^{10d}$ nodes are sufficient to approximate continuous functions, where δ depends on the continuity properties of the target function. By contrast, nearly the same proof yields a bound of $1 / \delta^{2d}$ for shallow ReLU networks; this gap suggests a tantalizing direction for future work, separating shallow ReLU networks and their linearization.

Learning to Move with Affordance Maps

- William Qi, Ravi Teja Mullapudi, Saurabh Gupta, Deva Ramanan
- abstract@open-review(Poster): The ability to autonomously explore and navigate a physical space is a fundamental requirement for virtually any mobile autonomous agent, from household robotic vacuums to autonomous vehicles. Traditional SLAM-based approaches for exploration and navigation largely focus on leveraging scene geometry, but fail to model dynamic objects (such as other agents) or semantic constraints (such as wet floors or doorways). Learning-based RL agents are an attractive alternative because they can incorporate both semantic and geometric information, but are notoriously sample inefficient, difficult to generalize to novel settings, and are difficult to interpret. In this paper, we combine the best of both worlds with a modular approach that {\em learns} a spatial representation of a scene that is trained to be effective when coupled with traditional geometric planners. Specifically, we design an agent that learns to predict a spatial affordance map that elucidates what parts of a scene are navigable through active self-supervised experience gathering. In contrast to most simulation environments that assume a static world, we evaluate our approach in the VizDoom simulator, using large-scale randomly-generated maps containing a variety of dynamic actors and hazards. We show that learned affordance maps can be used to augment traditional approaches for both exploration and navigation, providing significant improvements in performance.

Deep Learning For Symbolic Mathematics

- Guillaume Lample, François Charton
- abstract@open-review(Spotlight): Neural networks have a reputation for being better at solving statistical or approximate problems than at performing calculations or working with symbolic data. In this paper, we show that they can be surprisingly good at more elaborated tasks in mathematics, such as symbolic integration and solving differential equations. We propose a syntax for representing these mathematical problems, and methods for generating large datasets that can be used to train sequence-to-sequence models. We achieve results that outperform commercial Computer Algebra Systems such as Matlab or Mathematica.

Differentiable learning of numerical rules in knowledge graphs

- Po-Wei Wang, Daria Stepanova, Csaba Domokos, J. Zico Kolter
- abstract@open-review(Poster): Rules over a knowledge graph (KG) capture interpretable patterns in data and can be used for KG cleaning and completion. Inspired by the TensorLog differentiable logic framework, which compiles rule inference into a sequence of differentiable operations, recently a method called Neural LP has been proposed for learning the parameters as well as the structure of rules. However, it is limited with respect to the treatment of numerical features like age, weight or scientific measurements. We address this limitation by extending Neural LP to learn rules with numerical values, e.g., ”People younger than 18 typically live with their parents“. We demonstrate how dynamic programming and cumulative sum operations can be exploited to ensure efficiency of such extension. Our novel approach allows us to extract more expressive rules with aggregates, which are of higher quality and yield more accurate predictions compared to rules learned by the state-of-the-art methods, as shown by our experiments on synthetic and real-world datasets.

Consistency Regularization for Generative Adversarial Networks

- Han Zhang, Zizhao Zhang, Augustus Odena, Honglak Lee
- abstract@open-review(Poster): Generative Adversarial Networks (GANs) are known to be difficult to train, despite considerable research effort. Several regularization techniques for stabilizing training have been proposed, but they introduce non-trivial computational overheads and interact poorly with existing techniques like spectral normalization. In this work, we propose a simple, effective training stabilizer based on the notion of consistency regularization—a popular technique in the semi-supervised learning literature. In particular, we augment data passing into the GAN discriminator and penalize the sensitivity of the discriminator to these augmentations. We conduct a series of experiments to demonstrate that consistency regularization works effectively with spectral normalization and various GAN architectures, loss functions and optimizer settings. Our method achieves the best FID scores for unconditional image generation compared to other regularization methods on CIFAR-10 and CelebA. Moreover, Our consistency regularized GAN (CR-GAN) improves state of-the-art FID scores for conditional generation from 14.73 to 11.48 on CIFAR-10 and from 8.73 to 6.66 on ImageNet-2012.

On Generalization Error Bounds of Noisy Gradient Methods for Non-Convex Learning

- Jian Li, Xuanyuan Luo, Mingda Qiao
- abstract@open-review(Poster): Generalization error (also known as the out-of-sample error) measures how well the hypothesis learned from training data generalizes to previously unseen data. Proving tight generalization error bounds is a central question in statistical learning theory. In this paper, we obtain generalization error bounds for learning general non-convex objectives, which has attracted significant attention in recent years. We develop a new framework, termed Bayes-Stability, for proving algorithm-dependent generalization error bounds. The new framework combines ideas from both the PAC-Bayesian theory and the notion of algorithmic stability. Applying the Bayes-Stability method, we obtain new data-dependent generalization bounds for stochastic gradient Langevin dynamics (SGLD) and several other noisy gradient methods (e.g., with momentum, mini-batch and acceleration, Entropy-SGD). Our result recovers (and is typically tighter than) a recent result in Mou et al. (2018) and improves upon the results in Pensia et al. (2018). Our experiments demonstrate that our data-dependent bounds can distinguish randomly labelled data from normal data, which provides an explanation to the intriguing phenomena observed in Zhang et al. (2017a). We also study the setting where the total loss is the sum of a bounded loss and an additiona l`2 regularization term. We obtain new generalization bounds for the continuous Langevin dynamic in this setting by developing a new Log-Sobolev inequality for the parameter distribution at any time. Our new bounds are more desirable when the noise level of the processis not very small, and do not become vacuous even when T tends to infinity.

SUMO: Unbiased Estimation of Log Marginal Probability for Latent Variable Models

- Yucen Luo, Alex Beatson, Mohammad Norouzi, Jun Zhu, David Duvenaud, Ryan P. Adams, Ricky T. Q. Chen
- abstract@open-review(Spotlight): Standard variational lower bounds used to train latent variable models produce biased estimates of most quantities of interest. We introduce an unbiased estimator of the log marginal likelihood and its gradients for latent variable models based on randomized truncation of infinite series. If parameterized by an encoder-decoder architecture, the parameters of the encoder can be optimized to minimize its variance of this estimator. We show that models trained using our estimator give better test-set likelihoods than a standard importance-sampling based approach for the

same average computational cost. This estimator also allows use of latent variable models for tasks where unbiased estimators, rather than marginal likelihood lower bounds, are preferred, such as minimizing reverse KL divergences and estimating score functions.

Scale-Equivariant Steerable Networks

- Ivan Sosnovik, Michał Szmaja, Arnold Smeulders
- abstract@open-review(Poster): The effectiveness of Convolutional Neural Networks (CNNs) has been substantially attributed to their built-in property of translation equivariance. However, CNNs do not have embedded mechanisms to handle other types of transformations. In this work, we pay attention to scale changes, which regularly appear in various tasks due to the changing distances between the objects and the camera. First, we introduce the general theory for building scale-equivariant convolutional networks with steerable filters. We develop scale-convolution and generalize other common blocks to be scale-equivariant. We demonstrate the computational efficiency and numerical stability of the proposed method. We compare the proposed models to the previously developed methods for scale equivariance and local scale invariance. We demonstrate state-of-the-art results on the MNIST-scale dataset and on the STL-10 dataset in the supervised learning setting.

DeepSphere: a graph-based spherical CNN

- Michaël Defferrard, Martino Milani, Frédérick Gusset, Nathanaël Perraudin
- abstract@open-review(Spotlight): Designing a convolution for a spherical neural network requires a delicate tradeoff between efficiency and rotation equivariance. DeepSphere, a method based on a graph representation of the discretized sphere, strikes a controllable balance between these two desiderata. This contribution is twofold. First, we study both theoretically and empirically how equivariance is affected by the underlying graph with respect to the number of pixels and neighbors. Second, we evaluate DeepSphere on relevant problems. Experiments show state-of-the-art performance and demonstrates the efficiency and flexibility of this formulation. Perhaps surprisingly, comparison with previous work suggests that anisotropic filters might be an unnecessary price to pay. Our code is available at https://github.com/deepsphere.

Classification-Based Anomaly Detection for General Data

- Liron Bergman, Yedid Hoshen
- abstract@open-review(Poster): Anomaly detection, finding patterns that substantially deviate from those seen previously, is one of the fundamental problems of artificial intelligence. Recently, classification-based methods were shown to achieve superior results on this task. In this work, we present a unifying view and propose an open-set method, GOAD, to relax current generalization assumptions. Furthermore, we extend the applicability of transformation-based methods to non-image data using random affine transformations. Our method is shown to obtain state-of-the-art accuracy and is applicable to broad data types. The strong performance of our method is extensively validated on multiple datasets from different domains.

Unrestricted Adversarial Examples via Semantic Manipulation

- Anand Bhattad, Min Jin Chong, Kaizhao Liang, Bo Li, D. A. Forsyth
- abstract@open-review(Poster): Machine learning models, especially deep neural networks (DNNs), have been shown to be vulnerable against adversarial examples which are carefully crafted samples with a small magnitude of the perturbation. Such adversarial perturbations are usually restricted by bounding their $\mathcal{L}_p$ norm such that they are imperceptible, and thus many current defenses can exploit this property to reduce their adversarial impact. In this paper, we instead introduce "unrestricted" perturbations that manipulate semantically meaningful image-based visual descriptors - color and texture - in order to generate effective and photorealistic adversarial examples. We show that these semantically aware perturbations are effective against JPEG compression, feature squeezing and adversarially trained model. We also show that the proposed methods can effectively be applied to both image classification and image captioning tasks on complex datasets such as ImageNet and MSCOCO. In addition, we conduct comprehensive user studies to show that our generated semantic adversarial examples are photorealistic to humans despite large magnitude perturbations when compared to other attacks.

Discriminative Particle Filter Reinforcement Learning for Complex Partial observations

- Xiao Ma, Peter Karkus, David Hsu, Wee Sun Lee, Nan Ye
- abstract@open-review(Poster): Deep reinforcement learning is successful in decision making for sophisticated games, such as Atari, Go, etc. However, real-world decision making often requires reasoning with partial information extracted from complex visual observations. This paper presents Discriminative Particle Filter Reinforcement Learning (DPFRL), a new reinforcement learning framework for complex partial observations. DPFRL encodes a differentiable particle filter in the neural network policy for explicit reasoning with partial observations over time. The particle filter maintains a belief using learned discriminative update, which is trained end-to-end for decision making. We show that using the discriminative update instead of standard generative models results in significantly improved performance, especially for tasks with complex visual observations, because they circumvent the difficulty of modeling complex observations that are irrelevant to decision making. In addition, to extract features from the particle belief, we propose a new type of belief feature based on the moment generating function. DPFRL outperforms state-of-the-art POMDP RL models in Flickering Atari Games, an existing POMDP RL benchmark, and in Natural Flickering Atari Games, a new, more challenging POMDP RL benchmark introduced in this paper. Further, DPFRL performs well for visual navigation with real-world data in the Habitat environment.

Learning to Group: A Bottom-Up Framework for 3D Part Discovery in Unseen Categories

- Tiange Luo, Kaichun Mo, Zhiao Huang, Jiarui Xu, Siyu Hu, Liwei Wang, Hao Su
- abstract@open-review(Poster): We address the problem of learning to discover 3D parts for objects in unseen categories. Being able to learn the geometry prior of parts and transfer this prior to unseen categories pose fundamental challenges on data-driven shape segmentation approaches. Formulated as a contextual bandit problem, we propose a learning-based iterative grouping framework which learns a grouping policy to progressively merge small part proposals into bigger ones in a bottom-up fashion. At the core of our approach is to restrict the local context for extracting part-level features, which encourages the generalizability to novel categories. On a recently proposed large-scale fine-grained 3D part dataset, PartNet, we demonstrate that our method can transfer knowledge of parts learned from 3 training categories to 21 unseen testing categories without seeing any annotated samples. Quantitative comparisons against four strong shape segmentation baselines show that we achieve the state-of-the-art performance.

State Alignment-based Imitation Learning

- Fangchen Liu, Zhan Ling, Tongzhou Mu, Hao Su
- abstract@open-review(Poster): Consider an imitation learning problem that the imitator and the expert have different dynamics models. Most of existing imitation learning methods fail because they focus on the imitation of actions. We propose a novel state alignment-based imitation learning method to train the imitator by following the state sequences in the expert demonstrations as much as possible. The alignment of states comes from both local and global perspectives. We combine them into a reinforcement learning framework by a regularized policy update objective. We show the superiority of our method on standard imitation learning settings as well as the challenging settings in which the expert and the imitator have different dynamics models.

Neural Arithmetic Units

- Andreas Madsen, Alexander Rosenberg Johansen
- abstract@open-review(Spotlight): Neural networks can approximate complex functions, but they struggle to perform exact arithmetic operations over real numbers. The lack of inductive bias for arithmetic operations leaves neural networks without the underlying logic necessary to extrapolate on tasks such as addition, subtraction, and multiplication. We present two new neural network components: the Neural Addition Unit (NAU), which can learn exact addition and subtraction; and the Neural Multiplication Unit (NMU) that can multiply subsets of a vector. The NMU is, to our knowledge, the first arithmetic neural network component that can learn to multiply elements from a vector, when the hidden size is large. The two new components draw inspiration from a theoretical analysis of recently proposed arithmetic components. We find that careful initialization, restricting parameter space, and regularizing for sparsity is important when optimizing the NAU and NMU. Our proposed units NAU and NMU, compared with previous neural units, converge more consistently, have fewer parameters, learn faster, can converge for larger hidden sizes, obtain sparse and meaningful weights, and can extrapolate to negative and small values.

Lipschitz constant estimation of Neural Networks via sparse polynomial optimization

- Fabian Latorre, Paul Rolland, Volkan Cevher
- abstract@open-review(Poster): We introduce LiPopt, a polynomial optimization framework for computing increasingly tighter upper bound on the Lipschitz constant of neural networks. The underlying optimization problems boil down to either linear (LP) or semidefinite (SDP) programming. We show how to use the sparse connectivity of a network, to significantly reduce the complexity of computation. This is specially useful for convolutional as well as pruned neural networks. We conduct experiments on networks with random weights as well as networks trained on MNIST, showing that in the particular case of the ℓ_∞-Lipschitz constant, our approach yields superior estimates as compared to other baselines available in the literature.

Effect of Activation Functions on the Training of Overparametrized Neural Nets

- Abhishek Panigrahi, Abhishek Shetty, Navin Goyal
- abstract@open-review(Poster): It is well-known that overparametrized neural networks trained using gradient based methods quickly achieve small training error with appropriate hyperparameter settings. Recent papers have proved this statement theoretically for highly overparametrized networks under reasonable assumptions. These results either assume that the activation function is ReLU or they depend on the minimum eigenvalue of a certain Gram matrix. In the latter case, existing works only prove that this minimum eigenvalue is non-zero and do not provide quantitative bounds which require that this eigenvalue be large. Empirically, a number of alternative activation functions have been proposed which tend to perform better than ReLU at least in some settings but no clear understanding has emerged. This state of affairs underscores the importance of theoretically understanding the impact of activation functions on training. In the present paper, we provide theoretical results about the effect of activation function on the training of highly overparametrized 2-layer neural networks. A crucial property that governs the performance of an activation is whether or not it is smooth: • For non-smooth activations such as ReLU, SELU, ELU, which are not smooth because there is a point where either the first order or second order derivative is discontinuous, all eigenvalues of the associated Gram matrix are large under minimal assumptions on the data. • For smooth activations such as tanh, swish, polynomial, which have derivatives of all orders at all points, the situation is more complex: if the subspace spanned by the data has small dimension then the minimum eigenvalue of the Gram matrix can be small leading to slow training. But if the dimension is large and the data satisfies another mild condition, then the eigenvalues are large. If we allow deep networks, then the small data dimension is not a limitation provided that the depth is sufficient. We discuss a number of extensions and applications of these results.

Provable Filter Pruning for Efficient Neural Networks

- Lucas Liebenwein, Cenk Baykal, Harry Lang, Dan Feldman, Daniela Rus
- abstract@open-review(Poster): We present a provable, sampling-based approach for generating compact Convolutional Neural Networks (CNNs) by identifying and removing redundant filters from an over-parameterized network. Our algorithm uses a small batch of input data points to assign a saliency score to each filter and constructs an importance sampling distribution where filters that highly affect the output are sampled with correspondingly high probability. In contrast to existing filter pruning approaches, our method is simultaneously data-informed, exhibits provable guarantees on the size and performance of the pruned network, and is widely applicable to varying network architectures and data sets. Our analytical bounds bridge the notions of compressibility and importance of network structures, which gives rise to a fully-automated procedure for identifying and preserving filters in layers that are essential to the network's performance. Our experimental evaluations on popular architectures and data sets show that our algorithm consistently generates sparser and more efficient models than those constructed by existing filter pruning approaches.

End to End Trainable Active Contours via Differentiable Rendering

- Shir Gur, Tal Shaharabany, Lior Wolf
- abstract@open-review(Poster): We present an image segmentation method that iteratively evolves a polygon. At each iteration, the vertices of the polygon are displaced based on the local value of a 2D shift map that is inferred from the input image via an encoder-decoder architecture. The main training loss that is used is the difference between the polygon shape and the ground truth segmentation mask. The network employs a neural renderer to create the polygon from its vertices, making the process fully differentiable. We demonstrate that our method outperforms the state of the art segmentation networks and deep active contour solutions in a variety of benchmarks, including medical imaging and aerial images.

Compositional Language Continual Learning

- Yuanpeng Li, Liang Zhao, Kenneth Church, Mohamed Elhoseiny
- abstract@open-review(Poster): Motivated by the human's ability to continually learn and gain knowledge over time, several research efforts have been pushing the limits of machines to constantly learn while alleviating catastrophic forgetting. Most of the existing methods have been focusing on continual learning of label prediction tasks, which have fixed input and output sizes. In this paper, we propose a new scenario of continual learning which handles sequence-to-sequence tasks common in language learning. We further propose an approach to use label prediction continual learning algorithm for sequence-to-sequence continual learning by leveraging compositionality. Experimental results show that the proposed method has significant improvement over state-of-the-art methods. It enables knowledge transfer and prevents catastrophic forgetting, resulting in more than 85% accuracy up to 100 stages, compared with less than 50% accuracy for baselines in instruction learning task. It also shows significant improvement in machine translation task. This is the first work to combine continual learning and compositionality for language learning, and we hope this work will make machines more helpful in various tasks.

Adversarial Lipschitz Regularization

- Dávid Terjék
- abstract@open-review(Poster): Generative adversarial networks (GANs) are one of the most popular approaches when it comes to training generative models, among which variants of Wasserstein GANs are considered superior to the standard GAN formulation in terms of learning stability and sample quality. However, Wasserstein GANs require the critic to be 1-Lipschitz, which is often enforced implicitly by penalizing the norm of its gradient, or by globally restricting its Lipschitz constant via weight normalization techniques. Training with a regularization term penalizing the violation of the Lipschitz constraint explicitly, instead of through the norm of the gradient, was found to be practically infeasible in most situations. Inspired by Virtual Adversarial Training, we propose a method called Adversarial Lipschitz Regularization, and show that using an explicit Lipschitz penalty is indeed viable and leads to

competitive performance when applied to Wasserstein GANs, highlighting an important connection between Lipschitz regularization and adversarial training.

Adversarial Training and Provable Defenses: Bridging the Gap

- Mislav Balunovic, Martin Vechev
- abstract@open-review(Talk): We present COLT, a new method to train neural networks based on a novel combination of adversarial training and provable defenses. The key idea is to model neural network training as a procedure which includes both, the verifier and the adversary. In every iteration, the verifier aims to certify the network using convex relaxation while the adversary tries to find inputs inside that convex relaxation which cause verification to fail. We experimentally show that this training method, named convex layerwise adversarial training (COLT), is promising and achieves the best of both worlds -- it produces a state-of-the-art neural network with certified robustness of 60.5% and accuracy of 78.4% on the challenging CIFAR-10 dataset with a 2/255 L-infinity perturbation. This significantly improves over the best concurrent results of 54.0% certified robustness and 71.5% accuracy.

A Function Space View of Bounded Norm Infinite Width ReLU Nets: The Multivariate Case

- Greg Ongie, Rebecca Willett, Daniel Soudry, Nathan Srebro
- abstract@open-review(Poster): We give a tight characterization of the (vectorized Euclidean) norm of weights required to realize a function $f:\mathbb{R}\rightarrow \mathbb{R}^d$ as a single hidden-layer ReLU network with an unbounded number of units (infinite width), extending the univariate characterization of Savarese et al. (2019) to the multivariate case.

Lite Transformer with Long-Short Range Attention

- Zhanghao Wu, *Zhijian Liu*, Ji Lin, Yujun Lin, Song Han
- abstract@open-review(Poster): Transformer has become ubiquitous in natural language processing (e.g., machine translation, question answering); however, it requires enormous amount of computations to achieve high performance, which makes it not suitable for mobile applications that are tightly constrained by the hardware resources and battery. In this paper, we present an efficient mobile NLP architecture, Lite Transformer to facilitate deploying mobile NLP applications on edge devices. The key primitive is the Long-Short Range Attention (LSRA), where one group of heads specializes in the local context modeling (by convolution) while another group specializes in the long-distance relationship modeling (by attention). Such specialization brings consistent improvement over the vanilla transformer on three well-established language tasks: machine translation, abstractive summarization, and language modeling. Under constrained resources (500M/100M MACs), Lite Transformer outperforms transformer on WMT'14 English-French by 1.2/1.7 BLEU, respectively. Lite Transformer reduces the computation of transformer base model by 2.5x with 0.3 BLEU score degradation. Combining with pruning and quantization, we further compressed the model size of Lite Transformer by 18.2x. For language modeling, Lite Transformer achieves 1.8 lower perplexity than the transformer at around 500M MACs. Notably, Lite Transformer outperforms the AutoML-based Evolved Transformer by 0.5 higher BLEU for the mobile NLP setting without the costly architecture search that requires more than 250 GPU years. Code has been made available at https://github.com/mit-han-lab/lite-transformer.

Mutual Information Gradient Estimation for Representation Learning

- Liangjian Wen, Yiji Zhou, Lirong He, Mingyuan Zhou, Zenglin Xu
- abstract@open-review(Poster): Mutual Information (MI) plays an important role in representation learning. However, MI is unfortunately intractable in continuous and high-dimensional settings. Recent advances establish tractable and scalable MI estimators to discover useful representation. However, most of the existing methods are not capable of providing an accurate estimation of MI with low-variance when the MI is large. We argue that directly estimating the gradients of MI is more appealing for representation learning than estimating MI in itself. To this end, we propose the Mutual Information Gradient Estimator (MIGE) for representation learning based on the score estimation of implicit distributions. MIGE exhibits a tight and smooth gradient estimation of MI in the high-dimensional and large-MI settings. We expand the applications of MIGE in both unsupervised learning of deep representations based on InfoMax and the Information Bottleneck method. Experimental results have indicated significant performance improvement in learning useful representation.

Regularizing activations in neural networks via distribution matching with the Wasserstein metric

- Taejong Joo, Donggu Kang, Byunghoon Kim
- abstract@open-review(Poster): Regularization and normalization have become indispensable components in training deep neural networks, resulting in faster training and improved generalization performance. We propose the projected error function regularization loss (PER) that encourages activations to follow the standard normal distribution. PER randomly projects activations onto one-dimensional space and computes the regularization loss in the projected space. PER is similar to the Pseudo-Huber loss in the projected space, thus taking advantage of both L^1 and L^2 regularization losses. Besides, PER can capture the interaction between hidden units by projection vector drawn from a unit sphere. By doing so, PER minimizes the upper bound of the Wasserstein distance of order one between an empirical distribution of activations and the standard normal distribution. To the best of the authors' knowledge, this is the first work to regularize activations via distribution matching in the probability distribution space. We evaluate the proposed method on the image classification task and the word-level language modeling task.

Gradient Descent Maximizes the Margin of Homogeneous Neural Networks

- Kaifeng Lyu, Jian Li
- abstract@open-review(Talk): In this paper, we study the implicit regularization of the gradient descent algorithm in homogeneous neural networks, including fully-connected and convolutional neural networks with ReLU or LeakyReLU activations. In particular, we study the gradient descent or gradient flow (i.e., gradient descent with infinitesimal step size) optimizing the logistic loss or cross-entropy loss of any homogeneous model (possibly non-smooth), and show that if the training loss decreases below a certain threshold, then we can define a smoothed version of the normalized margin which increases over time. We also formulate a natural constrained optimization problem related to margin maximization, and prove that both the normalized margin and its smoothed version converge to the objective value at a KKT point of the optimization problem. Our results generalize the previous results for logistic regression with one-layer or multi-layer linear networks, and provide more quantitative convergence results with weaker assumptions than previous results for homogeneous smooth neural networks. We conduct several experiments to justify our theoretical finding on MNIST and CIFAR-10 datasets. Finally, as margin is closely related to robustness, we discuss potential benefits of training longer for improving the robustness of the model.

Transferring Optimality Across Data Distributions via Homotopy Methods

- Matilde Gargiani, Andrea Zanelli, Quoc Tran Dinh, Moritz Diehl, Frank Hutter
- abstract@open-review(Poster): Homotopy methods, also known as continuation methods, are a powerful mathematical tool to efficiently solve various problems in numerical analysis, including complex non-convex optimization problems where no or only little prior knowledge regarding the localization of the solutions is available. In this work, we propose a novel homotopy-based numerical method that can be used to transfer knowledge regarding the localization of an optimum across different task distributions in deep learning applications. We validate the proposed methodology with some empirical

evaluations in the regression and classification scenarios, where it shows that superior numerical performance can be achieved in popular deep learning benchmarks, i.e. FashionMNIST, CIFAR-10, and draw connections with the widely used fine-tuning heuristic. In addition, we give more insights on the properties of a general homotopy method when used in combination with Stochastic Gradient Descent by conducting a general local theoretical analysis in a simplified setting.

Latent Normalizing Flows for Many-to-Many Cross-Domain Mappings

- Shweta Mahajan, Iryna Gurevych, Stefan Roth
- abstract@open-review(Poster): Learned joint representations of images and text form the backbone of several important cross-domain tasks such as image captioning. Prior work mostly maps both domains into a common latent representation in a purely supervised fashion. This is rather restrictive, however, as the two domains follow distinct generative processes. Therefore, we propose a novel semi-supervised framework, which models shared information between domains and domain-specific information separately. The information shared between the domains is aligned with an invertible neural network. Our model integrates normalizing flow-based priors for the domain-specific information, which allows us to learn diverse many-to-many mappings between the two domains. We demonstrate the effectiveness of our model on diverse tasks, including image captioning and text-to-image synthesis.

Dynamic Model Pruning with Feedback

- Tao Lin, Sebastian U. Stich, Luis Barba, Daniil Dmitriev, Martin Jaggi
- abstract@open-review(Poster): Deep neural networks often have millions of parameters. This can hinder their deployment to low-end devices, not only due to high memory requirements but also because of increased latency at inference. We propose a novel model compression method that generates a sparse trained model without additional overhead: by allowing (i) dynamic allocation of the sparsity pattern and (ii) incorporating feedback signal to reactivate prematurely pruned weights we obtain a performant sparse model in one single training pass (retraining is not needed, but can further improve the performance). We evaluate the method on CIFAR-10 and ImageNet, and show that the obtained sparse models can reach the state-of-the-art performance of dense models and further that their performance surpasses all previously proposed pruning schemes (that come without feedback mechanisms).

On the interaction between supervision and self-play in emergent communication

- Ryan Lowe, *Abhinav Gupta*, Jakob Foerster, Douwe Kiela, Joelle Pineau
- abstract@open-review(Poster): A promising approach for teaching artificial agents to use natural language involves using human-in-the-loop training. However, recent work suggests that current machine learning methods are too data inefficient to be trained in this way from scratch. In this paper, we investigate the relationship between two categories of learning signals with the ultimate goal of improving sample efficiency: imitating human language data via supervised learning, and maximizing reward in a simulated multi-agent environment via self-play (as done in emergent communication), and introduce the term supervised self-play (S2P) for algorithms using both of these signals. We find that first training agents via supervised learning on human data followed by self-play outperforms the converse, suggesting that it is not beneficial to emerge languages from scratch. We then empirically investigate various S2P schedules that begin with supervised learning in two environments: a Lewis signaling game with symbolic inputs, and an image-based referential game with natural language descriptions. Lastly, we introduce population based approaches to S2P, which further improves the performance over single-agent methods.

A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms

- Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Nan Rosemary Ke, Sebastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, Christopher Pal
- abstract@open-review(Poster): We propose to use a meta-learning objective that maximizes the speed of transfer on a modified distribution to learn how to modularize acquired knowledge. In particular, we focus on how to factor a joint distribution into appropriate conditionals, consistent with the causal directions. We explain when this can work, using the assumption that the changes in distributions are localized (e.g. to one of the marginals, for example due to an intervention on one of the variables). We prove that under this assumption of localized changes in causal mechanisms, the correct causal graph will tend to have only a few of its parameters with non-zero gradient, i.e. that need to be adapted (those of the modified variables). We argue and observe experimentally that this leads to faster adaptation, and use this property to define a meta-learning surrogate score which, in addition to a continuous parametrization of graphs, would favour correct causal graphs. Finally, motivated by the AI agent point of view (e.g. of a robot discovering its environment autonomously), we consider how the same objective can discover the causal variables themselves, as a transformation of observed low-level variables with no causal meaning. Experiments in the two-variable case validate the proposed ideas and theoretical results.

Expected Information Maximization: Using the I-Projection for Mixture Density Estimation

- Philipp Becker, Oleg Arenz, Gerhard Neumann
- abstract@open-review(Poster): Modelling highly multi-modal data is a challenging problem in machine learning. Most algorithms are based on maximizing the likelihood, which corresponds to the M(oment)-projection of the data distribution to the model distribution. The M-projection forces the model to average over modes it cannot represent. In contrast, the I(nformation)-projection ignores such modes in the data and concentrates on the modes the model can represent. Such behavior is appealing whenever we deal with highly multi-modal data where modelling single modes correctly is more important than covering all the modes. Despite this advantage, the I-projection is rarely used in practice due to the lack of algorithms that can efficiently optimize it based on data. In this work, we present a new algorithm called Expected Information Maximization (EIM) for computing the I-projection solely based on samples for general latent variable models, where we focus on Gaussian mixtures models and Gaussian mixtures of experts. Our approach applies a variational upper bound to the I-projection objective which decomposes the original objective into single objectives for each mixture component as well as for the coefficients, allowing an efficient optimization. Similar to GANs, our approach employs discriminators but uses a more stable optimization procedure, using a tight upper bound. We show that our algorithm is much more effective in computing the I-projection than recent GAN approaches and we illustrate the effectiveness of our approach for modelling multi-modal behavior on two pedestrian and traffic prediction datasets.

Truth or backpropaganda? An empirical investigation of deep learning theory

- Micah Goldblum, Jonas Geiping, Avi Schwarzschild, Michael Moeller, Tom Goldstein
- abstract@open-review(Spotlight): We empirically evaluate common assumptions about neural networks that are widely held by practitioners and theorists alike. In this work, we: (1) prove the widespread existence of suboptimal local minima in the loss landscape of neural networks, and we use our theory to find examples; (2) show that small-norm parameters are not optimal for generalization; (3) demonstrate that ResNets do not conform to wide-network theories, such as the neural tangent kernel, and that the interaction between skip connections and batch normalization plays a role; (4) find that rank does not correlate with generalization or robustness in a practical setting.

Deep Audio Priors Emerge From Harmonic Convolutional Networks

- Zhoutong Zhang, Yunyun Wang, Chuang Gan, Jiajun Wu, Joshua B. Tenenbaum, Antonio Torralba, William T. Freeman
- abstract@open-review(Poster): Convolutional neural networks (CNNs) excel in image recognition and generation. Among many efforts to explain their effectiveness, experiments show that CNNs carry strong inductive biases that capture natural image priors. Do deep networks also have inductive biases

for audio signals? In this paper, we empirically show that current network architectures for audio processing do not show strong evidence in capturing such priors. We propose Harmonic Convolution, an operation that helps deep networks distill priors in audio signals by explicitly utilizing the harmonic structure within. This is done by engineering the kernel to be supported by sets of harmonic series, instead of local neighborhoods for convolutional kernels. We show that networks using Harmonic Convolution can reliably model audio priors and achieve high performance in unsupervised audio restoration tasks. With Harmonic Convolution, they also achieve better generalization performance for sound source separation.

Drawing Early-Bird Tickets: Toward More Efficient Training of Deep Networks

- Haoran You, Chaojian Li, Pengfei Xu, Yonggan Fu, Yue Wang, Xiaohan Chen, Richard G. Baraniuk, Zhangyang Wang, Yingyan Lin
- abstract@open-review(Spotlight): (Frankle & Carbin, 2019) shows that there exist winning tickets (small but critical subnetworks) for dense, randomly initialized networks, that can be trained alone to achieve comparable accuracies to the latter in a similar number of iterations. However, the identification of these winning tickets still requires the costly train-prune-retrain process, limiting their practical benefits. In this paper, we discover for the first time that the winning tickets can be identified at the very early training stage, which we term as Early-Bird (EB) tickets, via low-cost training schemes (e.g., early stopping and low-precision training) at large learning rates. Our finding of EB tickets is consistent with recently reported observations that the key connectivity patterns of neural networks emerge early. Furthermore, we propose a mask distance metric that can be used to identify EB tickets with low computational overhead, without needing to know the true winning tickets that emerge after the full training. Finally, we leverage the existence of EB tickets and the proposed mask distance to develop efficient training methods, which are achieved by first identifying EB tickets via low-cost schemes, and then continuing to train merely the EB tickets towards the target accuracy. Experiments based on various deep networks and datasets validate: 1) the existence of EB tickets and the effectiveness of mask distance in efficiently identifying them; and 2) that the proposed efficient training via EB tickets can achieve up to 5.8x ~ 10.7x energy savings while maintaining comparable or even better accuracy as compared to the most competitive state-of-the-art training methods, demonstrating a promising and easily adopted method for tackling cost-prohibitive deep network training.

Improving Generalization in Meta Reinforcement Learning using Learned Objectives

- Louis Kirsch, Sjoerd van Steenkiste, Juergen Schmidhuber
- abstract@open-review(Spotlight): Biological evolution has distilled the experiences of many learners into the general learning algorithms of humans. Our novel meta reinforcement learning algorithm MetaGenRL is inspired by this process. MetaGenRL distills the experiences of many complex agents to meta-learn a low-complexity neural objective function that decides how future individuals will learn. Unlike recent meta-RL algorithms, MetaGenRL can generalize to new environments that are entirely different from those used for meta-training. In some cases, it even outperforms human-engineered RL algorithms. MetaGenRL uses off-policy second-order gradients during meta-training that greatly increase its sample efficiency.

A closer look at the approximation capabilities of neural networks

- Kai Fong Ernest Chong
- abstract@open-review(Poster): The universal approximation theorem, in one of its most general versions, says that if we consider only continuous activation functions σ, then a standard feedforward neural network with one hidden layer is able to approximate any continuous multivariate function f to any given approximation threshold ε, if and only if σ is non-polynomial. In this paper, we give a direct algebraic proof of the theorem. Furthermore we shall explicitly quantify the number of hidden units required for approximation. Specifically, if X in R^n is compact, then a neural network with n input units, m output units, and a single hidden layer with {n+d choose d} hidden units (independent of m and ε), can uniformly approximate any polynomial function f:X -> R^m whose total degree is at most d for each of its m coordinate functions. In the general case that f is any continuous function, we show there exists some N in O(ε^{-n}) (independent of m), such that N hidden units would suffice to approximate f. We also show that this uniform approximation property (UAP) still holds even under seemingly strong conditions imposed on the weights. We highlight several consequences: (i) For any $\delta > 0$, the UAP still holds if we restrict all non-bias weights w in the last layer to satisfy $|w| < \delta$. (ii) There exists some $\lambda>0$ (depending only on f and σ), such that the UAP still holds if we restrict all non-bias weights w in the first layer to satisfy $|w|>\lambda$. (iii) If the non-bias weights in the first layer are *fixed* and randomly chosen from a suitable range, then the UAP holds with probability 1.

Kaleidoscope: An Efficient, Learnable Representation For All Structured Linear Maps

- Tri Dao, Nimit Sohoni, Albert Gu, Matthew Eichhorn, Amit Blonder, Megan Leszczynski, Atri Rudra, Christopher Ré
- abstract@open-review(Spotlight): Modern neural network architectures use structured linear transformations, such as low-rank matrices, sparse matrices, permutations, and the Fourier transform, to improve inference speed and reduce memory usage compared to general linear maps. However, choosing which of the myriad structured transformations to use (and its associated parameterization) is a laborious task that requires trading off speed, space, and accuracy. We consider a different approach: we introduce a family of matrices called kaleidoscope matrices (K-matrices) that provably capture any structured matrix with near-optimal space (parameter) and time (arithmetic operation) complexity. We empirically validate that K-matrices can be automatically learned within end-to-end pipelines to replace hand-crafted procedures, in order to improve model quality. For example, replacing channel shuffles in ShuffleNet improves classification accuracy on ImageNet by up to 5%. K-matrices can also simplify hand-engineered pipelines---we replace filter bank feature computation in speech data preprocessing with a learnable kaleidoscope layer, resulting in only 0.4% loss in accuracy on the TIMIT speech recognition task. In addition, K-matrices can capture latent structure in models: for a challenging permuted image classification task, adding a K-matrix to a standard convolutional architecture can enable learning the latent permutation and improve accuracy by over 8 points. We provide a practically efficient implementation of our approach, and use K-matrices in a Transformer network to attain 36% faster end-to-end inference speed on a language translation task.

Residual Energy-Based Models for Text Generation

- Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, Marc'Aurelio Ranzato
- abstract@open-review(Poster): Text generation is ubiquitous in many NLP tasks, from summarization, to dialogue and machine translation. The dominant parametric approach is based on locally normalized models which predict one word at a time. While these work remarkably well, they are plagued by exposure bias due to the greedy nature of the generation process. In this work, we investigate un-normalized energy-based models (EBMs) which operate not at the token but at the sequence level. In order to make training tractable, we first work in the residual of a pretrained locally normalized language model and second we train using noise contrastive estimation. Furthermore, since the EBM works at the sequence level, we can leverage pretrained bi-directional contextual representations, such as BERT and RoBERTa. Our experiments on two large language modeling datasets show that residual EBMs yield lower perplexity compared to locally normalized baselines. Moreover, generation via importance sampling is very efficient and of higher quality than the baseline models according to human evaluation.

AtomNAS: Fine-Grained End-to-End Neural Architecture Search

- Jieru Mei, Yingwei Li, Xiaochen Lian, Xiaojie Jin, Linjie Yang, Alan Yuille, Jianchao Yang
- abstract@open-review(Poster): Search space design is very critical to neural architecture search (NAS) algorithms. We propose a fine-grained search space comprised of atomic blocks, a minimal search unit that is much smaller than the ones used in recent NAS algorithms. This search space allows a mix of operations by composing different types of atomic blocks, while the search space in previous methods only allows homogeneous operations. Based on this search space, we propose a resource-aware architecture search framework which automatically assigns the computational resources (e.g., output channel numbers) for each operation by jointly considering the performance and the computational cost. In addition, to accelerate the search process, we

propose a dynamic network shrinkage technique which prunes the atomic blocks with negligible influence on outputs on the fly. Instead of a search-and-retrain two-stage paradigm, our method simultaneously searches and trains the target architecture. Our method achieves state-of-the-art performance under several FLOPs configurations on ImageNet with a small searching cost. We open our entire codebase at: https://github.com/meijieru/AtomNAS.

AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty

- Dan Hendrycks, *Norman Mu*, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, Balaji Lakshminarayanan
- abstract@open-review(Poster): Modern deep neural networks can achieve high accuracy when the training distribution and test distribution are identically distributed, but this assumption is frequently violated in practice. When the train and test distributions are mismatched, accuracy can plummet. Currently there are few techniques that improve robustness to unforeseen data shifts encountered during deployment. In this work, we propose a technique to improve the robustness and uncertainty estimates of image classifiers. We propose AugMix, a data processing technique that is simple to implement, adds limited computational overhead, and helps models withstand unforeseen corruptions. AugMix significantly improves robustness and uncertainty measures on challenging image classification benchmarks, closing the gap between previous methods and the best possible performance in some cases by more than half.

Disentanglement by Nonlinear ICA with General Incompressible-flow Networks (GIN)

- Peter Sorrenson, Carsten Rother, Ullrich Köthe
- abstract@open-review(Spotlight): A central question of representation learning asks under which conditions it is possible to reconstruct the true latent variables of an arbitrarily complex generative process. Recent breakthrough work by Khemakhem et al. (2019) on nonlinear ICA has answered this question for a broad class of conditional generative processes. We extend this important result in a direction relevant for application to real-world data. First, we generalize the theory to the case of unknown intrinsic problem dimension and prove that in some special (but not very restrictive) cases, informative latent variables will be automatically separated from noise by an estimating model. Furthermore, the recovered informative latent variables will be in one-to-one correspondence with the true latent variables of the generating process, up to a trivial component-wise transformation. Second, we introduce a modification of the RealNVP invertible neural network architecture (Dinh et al. (2016)) which is particularly suitable for this type of problem: the General Incompressible-flow Network (GIN). Experiments on artificial data and EMNIST demonstrate that theoretical predictions are indeed verified in practice. In particular, we provide a detailed set of exactly 22 informative latent variables extracted from EMNIST.

Memory-Based Graph Networks

- Amir Hosein Khasahmadi, Kaveh Hassani, Parsa Moradi, Leo Lee, Quaid Morris
- abstract@open-review(Poster): Graph neural networks (GNNs) are a class of deep models that operate on data with arbitrary topology represented as graphs. We introduce an efficient memory layer for GNNs that can jointly learn node representations and coarsen the graph. We also introduce two new networks based on this layer: memory-based GNN (MemGNN) and graph memory network (GMN) that can learn hierarchical graph representations. The experimental results shows that the proposed models achieve state-of-the-art results in eight out of nine graph classification and regression benchmarks. We also show that the learned representations could correspond to chemical features in the molecule data.

Variational Template Machine for Data-to-Text Generation

- Rong Ye, Wenxian Shi, Hao Zhou, Zhongyu Wei, Lei Li
- abstract@open-review(Poster): How to generate descriptions from structured data organized in tables? Existing approaches using neural encoder-decoder models often suffer from lacking diversity. We claim that an open set of templates is crucial for enriching the phrase constructions and realizing varied generations.Learning such templates is prohibitive since it often requires a large paired , which is seldom available. This paper explores the problem of automatically learning reusable "templates" from paired and non-paired data. We propose the variational template machine (VTM), a novel method to generate text descriptions from data tables. Our contributions include: a) we carefully devise a specific model architecture and losses to explicitly disentangle text template and semantic content information, in the latent spaces, and b) we utilize both small parallel data and large raw text without aligned tables to enrich the template learning. Experiments on datasets from a variety of different domains show that VTM is able to generate more diversely while keeping a good fluency and quality.

Phase Transitions for the Information Bottleneck in Representation Learning

- Tailin Wu, Ian Fischer
- abstract@open-review(Poster): In the Information Bottleneck (IB), when tuning the relative strength between compression and prediction terms, how do the two terms behave, and what's their relationship with the dataset and the learned representation? In this paper, we set out to answer these questions by studying multiple phase transitions in the IB objective: $IB_\beta[p(z|x)] = I(X; Z) - \beta I(Y; Z)$ defined on the encoding distribution $p(z|x)$ for input X, target Y and representation Z, where sudden jumps of $dI(Y; Z)/d\beta$ and prediction accuracy are observed with increasing β. We introduce a definition for IB phase transitions as a qualitative change of the IB loss landscape, and show that the transitions correspond to the onset of learning new classes. Using second-order calculus of variations, we derive a formula that provides a practical condition for IB phase transitions, and draw its connection with the Fisher information matrix for parameterized models. We provide two perspectives to understand the formula, revealing that each IB phase transition is finding a component of maximum (nonlinear) correlation between X and Y orthogonal to the learned representation, in close analogy with canonical-correlation analysis (CCA) in linear settings. Based on the theory, we present an algorithm for discovering phase transition points. Finally, we verify that our theory and algorithm accurately predict phase transitions in categorical datasets, predict the onset of learning new classes and class difficulty in MNIST, and predict prominent phase transitions in CIFAR10.

Continual learning with hypernetworks

- Johannes von Oswald, Christian Henning, Benjamin F. Grewe, João Sacramento
- abstract@open-review(Spotlight): Artificial neural networks suffer from catastrophic forgetting when they are sequentially trained on multiple tasks. To overcome this problem, we present a novel approach based on task-conditioned hypernetworks, i.e., networks that generate the weights of a target model based on task identity. Continual learning (CL) is less difficult for this class of models thanks to a simple key feature: instead of recalling the input-output relations of all previously seen data, task-conditioned hypernetworks only require rehearsing task-specific weight realizations, which can be maintained in memory using a simple regularizer. Besides achieving state-of-the-art performance on standard CL benchmarks, additional experiments on long task sequences reveal that task-conditioned hypernetworks display a very large capacity to retain previous memories. Notably, such long memory lifetimes are achieved in a compressive regime, when the number of trainable hypernetwork weights is comparable or smaller than target network size. We provide insight into the structure of low-dimensional task embedding spaces (the input space of the hypernetwork) and show that task-conditioned hypernetworks demonstrate transfer learning. Finally, forward information transfer is further supported by empirical results on a challenging CL benchmark based on the CIFAR-10/100 image datasets.

Permutation Equivariant Models for Compositional Generalization in Language

- Jonathan Gordon, David Lopez-Paz, Marco Baroni, Diane Bouchacourt

- abstract@open-review(Poster): Humans understand novel sentences by composing meanings and roles of core language components. In contrast, neural network models for natural language modeling fail when such compositional generalization is required. The main contribution of this paper is to hypothesize that language compositionality is a form of group-equivariance. Based on this hypothesis, we propose a set of tools for constructing equivariant sequence-to-sequence models. Throughout a variety of experiments on the SCAN tasks, we analyze the behavior of existing models under the lens of equivariance, and demonstrate that our equivariant architecture is able to achieve the type compositional generalization required in human language understanding.

Training binary neural networks with real-to-binary convolutions

- Brais Martinez, Jing Yang, Adrian Bulat, Georgios Tzimiropoulos
- abstract@open-review(Poster): This paper shows how to train binary networks to within a few percent points (~3-5%) of the full precision counterpart. We first show how to build a strong baseline, which already achieves state-of-the-art accuracy, by combining recently proposed advances and carefully adjusting the optimization procedure. Secondly, we show that by attempting to minimize the discrepancy between the output of the binary and the corresponding real-valued convolution, additional significant accuracy gains can be obtained. We materialize this idea in two complementary ways: (1) with a loss function, during training, by matching the spatial attention maps computed at the output of the binary and real-valued convolutions, and (2) in a data-driven manner, by using the real-valued activations, available during inference prior to the binarization process, for re-scaling the activations right after the binary convolution. Finally, we show that, when putting all of our improvements together, the proposed model beats the current state of the art by more than 5% top-1 accuracy on ImageNet and reduces the gap to its real-valued counterpart to less than 3% and 5% top-1 accuracy on CIFAR-100 and ImageNet respectively when using a ResNet-18 architecture. Code available at https://github.com/brais-martinez/real2binary

StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding

- Wei Wang, Bin Bi, Ming Yan, Chen Wu, Jiangnan Xia, Zuyi Bao, Liwei Peng, Luo Si
- abstract@open-review(Poster): Recently, the pre-trained language model, BERT (and its robustly optimized version RoBERTa), has attracted a lot of attention in natural language understanding (NLU), and achieved state-of-the-art accuracy in various NLU tasks, such as sentiment classification, natural language inference, semantic textual similarity and question answering. Inspired by the linearization exploration work of Elman, we extend BERT to a new model, StructBERT, by incorporating language structures into pre-training. Specifically, we pre-train StructBERT with two auxiliary tasks to make the most of the sequential order of words and sentences, which leverage language structures at the word and sentence levels, respectively. As a result, the new model is adapted to different levels of language understanding required by downstream tasks.

The StructBERT with structural pre-training gives surprisingly good empirical results on a variety of downstream tasks, including pushing the state-of-the-art on the GLUE benchmark to 89.0 (outperforming all published models at the time of model submission), the F1 score on SQuAD v1.1 question answering to 93.0, the accuracy on SNLI to 91.7.

Smooth markets: A basic mechanism for organizing gradient-based learners

- David Balduzzi, Wojciech M. Czarnecki, Tom Anthony, Ian Gemp, Edward Hughes, Joel Leibo, Georgios Piliouras, Thore Graepel
- abstract@open-review(Poster): With the success of modern machine learning, it is becoming increasingly important to understand and control how learning algorithms interact. Unfortunately, negative results from game theory show there is little hope of understanding or controlling general n-player games. We therefore introduce smooth markets (SM-games), a class of n-player games with pairwise zero sum interactions. SM-games codify a common design pattern in machine learning that includes some GANs, adversarial training, and other recent algorithms. We show that SM-games are amenable to analysis and optimization using first-order methods.

Fair Resource Allocation in Federated Learning

- Tian Li, Maziar Sanjabi, Ahmad Beirami, Virginia Smith
- abstract@open-review(Poster): Federated learning involves training statistical models in massive, heterogeneous networks. Naively minimizing an aggregate loss function in such a network may disproportionately advantage or disadvantage some of the devices. In this work, we propose q-Fair Federated Learning (q-FFL), a novel optimization objective inspired by fair resource allocation in wireless networks that encourages a more fair (specifically, a more uniform) accuracy distribution across devices in federated networks. To solve q-FFL, we devise a communication-efficient method, q-FedAvg, that is suited to federated networks. We validate both the effectiveness of q-FFL and the efficiency of q-FedAvg on a suite of federated datasets with both convex and non-convex models, and show that q-FFL (along with q-FedAvg) outperforms existing baselines in terms of the resulting fairness, flexibility, and efficiency.

Convolutional Conditional Neural Processes

- Jonathan Gordon, Wessel P. Bruinsma, Andrew Y. K. Foong, James Requeima, Yann Dubois, Richard E. Turner
- abstract@open-review(Talk): We introduce the Convolutional Conditional Neural Process (ConvCNP), a new member of the Neural Process family that models translation equivariance in the data. Translation equivariance is an important inductive bias for many learning problems including time series modelling, spatial data, and images. The model embeds data sets into an infinite-dimensional function space, as opposed to finite-dimensional vector spaces. To formalize this notion, we extend the theory of neural representations of sets to include functional representations, and demonstrate that any translation-equivariant embedding can be represented using a convolutional deep-set. We evaluate ConvCNPs in several settings, demonstrating that they achieve state-of-the-art performance compared to existing NPs. We demonstrate that building in translation equivariance enables zero-shot generalization to challenging, out-of-domain tasks.

Meta-Learning with Warped Gradient Descent

- Sebastian Flennerhag, Andrei A. Rusu, Razvan Pascanu, Francesco Visin, Hujun Yin, Raia Hadsell
- abstract@open-review(Talk): Learning an efficient update rule from data that promotes rapid learning of new tasks from the same distribution remains an open problem in meta-learning. Typically, previous works have approached this issue either by attempting to train a neural network that directly produces updates or by attempting to learn better initialisations or scaling factors for a gradient-based update rule. Both of these approaches pose challenges. On one hand, directly producing an update forgoes a useful inductive bias and can easily lead to non-converging behaviour. On the other hand, approaches that try to control a gradient-based update rule typically resort to computing gradients through the learning process to obtain their meta-gradients, leading to methods that can not scale beyond few-shot task adaptation. In this work, we propose Warped Gradient Descent (WarpGrad), a method that intersects these approaches to mitigate their limitations. WarpGrad meta-learns an efficiently parameterised preconditioning matrix that facilitates gradient descent across the task distribution. Preconditioning arises by interleaving non-linear layers, referred to as warp-layers, between the layers of a task-learner. Warp-layers are meta-learned without backpropagating through the task training process in a manner similar to methods that learn to directly produce updates. WarpGrad is computationally efficient, easy to implement, and can scale to arbitrarily large meta-learning problems. We provide a geometrical interpretation of the approach and evaluate its effectiveness in a variety of settings, including few-shot, standard supervised, continual and reinforcement learning.

Never Give Up: Learning Directed Exploration Strategies

- Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martin Arjovsky, Alexander Pritzel, Andrew Bolt, Charles Blundell
- abstract@open-review(Poster): We propose a reinforcement learning agent to solve hard exploration games by learning a range of directed exploratory policies. We construct an episodic memory-based intrinsic reward using k-nearest neighbors over the agent's recent experience to train the directed exploratory policies, thereby encouraging the agent to repeatedly revisit all states in its environment. A self-supervised inverse dynamics model is used to train the embeddings of the nearest neighbour lookup, biasing the novelty signal towards what the agent can control. We employ the framework of Universal Value Function Approximators to simultaneously learn many directed exploration policies with the same neural network, with different trade-offs between exploration and exploitation. By using the same neural network for different degrees of exploration/exploitation, transfer is demonstrated from predominantly exploratory policies yielding effective exploitative policies. The proposed method can be incorporated to run with modern distributed RL agents that collect large amounts of experience from many actors running in parallel on separate environment instances. Our method doubles the performance of the base agent in all hard exploration in the Atari-57 suite while maintaining a very high score across the remaining games, obtaining a median human normalised score of 1344.0%. Notably, the proposed method is the first algorithm to achieve non-zero rewards (with a mean score of 8,400) in the game of Pitfall! without using demonstrations or hand-crafted features.

AdvectiveNet: An Eulerian-Lagrangian Fluidic Reservoir for Point Cloud Processing

- Xingzhe He, Helen Lu Cao, Bo Zhu
- abstract@open-review(Poster): This paper presents a novel physics-inspired deep learning approach for point cloud processing motivated by the natural flow phenomena in fluid mechanics. Our learning architecture jointly defines data in an Eulerian world space, using a static background grid, and a Lagrangian material space, using moving particles. By introducing this Eulerian-Lagrangian representation, we are able to naturally evolve and accumulate particle features using flow velocities generated from a generalized, high-dimensional force field. We demonstrate the efficacy of this system by solving various point cloud classification and segmentation problems with state-of-the-art performance. The entire geometric reservoir and data flow mimic the pipeline of the classic PIC/FLIP scheme in modeling natural flow, bridging the disciplines of geometric machine learning and physical simulation.

SEED RL: Scalable and Efficient Deep-RL with Accelerated Central Inference

- Lasse Espeholt, Raphaël Marinier, Piotr Stanczyk, Ke Wang, Marcin Michalski
- abstract@open-review(Talk): We present a modern scalable reinforcement learning agent called SEED (Scalable, Efficient Deep-RL). By effectively utilizing modern accelerators, we show that it is not only possible to train on millions of frames per second but also to lower the cost. of experiments compared to current methods. We achieve this with a simple architecture that features centralized inference and an optimized communication layer. SEED adopts two state-of-the-art distributed algorithms, IMPALA/V-trace (policy gradients) and R2D2 (Q-learning), and is evaluated on Atari-57, DeepMind Lab and Google Research Football. We improve the state of the art on Football and are able to reach state of the art on Atari-57 twice as fast in wall-time. For the scenarios we consider, a 40% to 80% cost reduction for running experiments is achieved. The implementation along with experiments is open-sourced so results can be reproduced and novel ideas tried out.

You CAN Teach an Old Dog New Tricks! On Training Knowledge Graph Embeddings

- Daniel Ruffinelli, Samuel Broscheit, Rainer Gemulla
- abstract@open-review(Poster): Knowledge graph embedding (KGE) models learn algebraic representations of the entities and relations in a knowledge graph. A vast number of KGE techniques for multi-relational link prediction have been proposed in the recent literature, often with state-of-the-art performance. These approaches differ along a number of dimensions, including different model architectures, different training strategies, and different approaches to hyperparameter optimization. In this paper, we take a step back and aim to summarize and quantify empirically the impact of each of these dimensions on model performance. We report on the results of an extensive experimental study with popular model architectures and training strategies across a wide range of hyperparameter settings. We found that when trained appropriately, the relative performance differences between various model architectures often shrinks and sometimes even reverses when compared to prior results. For example, RESCAL~\citep{nickel2011three}, one of the first KGE models, showed strong performance when trained with state-of-the-art techniques; it was competitive to or outperformed more recent architectures. We also found that good (and often superior to prior studies) model configurations can be found by exploring relatively few random samples from a large hyperparameter space. Our results suggest that many of the more advanced architectures and techniques proposed in the literature should be revisited to reassess their individual benefits. To foster further reproducible research, we provide all our implementations and experimental results as part of the open source LibKGE framework.

High Fidelity Speech Synthesis with Adversarial Networks

- Mikołaj Bińkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C. Cobo, Karen Simonyan
- abstract@open-review(Talk): Generative adversarial networks have seen rapid development in recent years and have led to remarkable improvements in generative modelling of images. However, their application in the audio domain has received limited attention, and autoregressive models, such as WaveNet, remain the state of the art in generative modelling of audio signals such as human speech. To address this paucity, we introduce GAN-TTS, a Generative Adversarial Network for Text-to-Speech. Our architecture is composed of a conditional feed-forward generator producing raw speech audio, and an ensemble of discriminators which operate on random windows of different sizes. The discriminators analyse the audio both in terms of general realism, as well as how well the audio corresponds to the utterance that should be pronounced. To measure the performance of GAN-TTS, we employ both subjective human evaluation (MOS - Mean Opinion Score), as well as novel quantitative metrics (Fréchet DeepSpeech Distance and Kernel DeepSpeech Distance), which we find to be well correlated with MOS. We show that GAN-TTS is capable of generating high-fidelity speech with naturalness comparable to the state-of-the-art models, and unlike autoregressive models, it is highly parallelisable thanks to an efficient feed-forward generator. Listen to GAN-TTS reading this abstract at https://storage.googleapis.com/deepmind-media/research/abstract.wav

At Stability's Edge: How to Adjust Hyperparameters to Preserve Minima Selection in Asynchronous Training of Neural Networks?

- Niv Giladi, Mor Shpigel Nacson, Elad Hoffer, Daniel Soudry
- abstract@open-review(Spotlight): Background: Recent developments have made it possible to accelerate neural networks training significantly using large batch sizes and data parallelism. Training in an asynchronous fashion, where delay occurs, can make training even more scalable. However, asynchronous training has its pitfalls, mainly a degradation in generalization, even after convergence of the algorithm. This gap remains not well understood, as theoretical analysis so far mainly focused on the convergence rate of asynchronous methods. Contributions: We examine asynchronous training from the perspective of dynamical stability. We find that the degree of delay interacts with the learning rate, to change the set of minima accessible by an asynchronous stochastic gradient descent algorithm. We derive closed-form rules on how the learning rate could be changed, while keeping the accessible set the same. Specifically, for high delay values, we find that the learning rate should be kept inversely proportional to the delay. We then extend this analysis to include momentum. We find momentum should be either turned off, or modified to improve training stability. We provide empirical experiments to validate our theoretical findings.

Functional Regularisation for Continual Learning with Gaussian Processes

- Michalis K. Titsias, Jonathan Schwarz, Alexander G. de G. Matthews, Razvan Pascanu, Yee Whye Teh
- abstract@open-review(Poster): We introduce a framework for Continual Learning (CL) based on Bayesian inference over the function space rather than the parameters of a deep neural network. This method, referred to as functional regularisation for Continual Learning, avoids forgetting a previous task by constructing and memorising an approximate posterior belief over the underlying task-specific function. To achieve this we rely on a Gaussian process obtained by treating the weights of the last layer of a neural network as random and Gaussian distributed. Then, the training algorithm sequentially encounters tasks and constructs posterior beliefs over the task-specific functions by using inducing point sparse Gaussian process methods. At each step a new task is first learnt and then a summary is constructed consisting of (i) inducing inputs – a fixed-size subset of the task inputs selected such that it optimally represents the task – and (ii) a posterior distribution over the function values at these inputs. This summary then regularises learning of future tasks, through Kullback-Leibler regularisation terms. Our method thus unites approaches focused on (pseudo-)rehearsal with those derived from a sequential Bayesian inference perspective in a principled way, leading to strong results on accepted benchmarks.

Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network

- Taiji Suzuki, Hiroshi Abe, Tomoaki Nishimura
- abstract@open-review(Spotlight): One of the biggest issues in deep learning theory is the generalization ability of networks with huge model size. The classical learning theory suggests that overparameterized models cause overfitting. However, practically used large deep models avoid overfitting, which is not well explained by the classical approaches. To resolve this issue, several attempts have been made. Among them, the compression based bound is one of the promising approaches. However, the compression based bound can be applied only to a compressed network, and it is not applicable to the non-compressed original network. In this paper, we give a unified frame-work that can convert compression based bounds to those for non-compressed original networks. The bound gives even better rate than the one for the compressed network by improving the bias term. By establishing the unified frame-work, we can obtain a data dependent generalization error bound which gives a tighter evaluation than the data independent ones.

Dynamics-Aware Embeddings

- William Whitney, Rajat Agarwal, Kyunghyun Cho, Abhinav Gupta
- abstract@open-review(Poster): In this paper we consider self-supervised representation learning to improve sample efficiency in reinforcement learning (RL). We propose a forward prediction objective for simultaneously learning embeddings of states and actions. These embeddings capture the structure of the environment's dynamics, enabling efficient policy learning. We demonstrate that our action embeddings alone improve the sample efficiency and peak performance of model-free RL on control from low-dimensional states. By combining state and action embeddings, we achieve efficient learning of high-quality policies on goal-conditioned continuous control from pixel observations in only 1-2 million environment steps.

RaPP: Novelty Detection with Reconstruction along Projection Pathway

- Ki Hyun Kim, Sangwoo Shim, Yongsub Lim, Jongseob Jeon, Jeongwoo Choi, Byungchan Kim, Andre S. Yoon
- abstract@open-review(Poster): We propose RaPP, a new methodology for novelty detection by utilizing hidden space activation values obtained from a deep autoencoder. Precisely, RaPP compares input and its autoencoder reconstruction not only in the input space but also in the hidden spaces. We show that if we feed a reconstructed input to the same autoencoder again, its activated values in a hidden space are equivalent to the corresponding reconstruction in that hidden space given the original input. In order to aggregate the hidden space activation values, we propose two metrics, which enhance the novelty detection performance. Through extensive experiments using diverse datasets, we validate that RaPP improves novelty detection performances of autoencoder-based approaches. Besides, we show that RaPP outperforms recent novelty detection methods evaluated on popular benchmarks.

Hypermodels for Exploration

- Vikranth Dwaracherla, Xiuyuan Lu, Morteza Ibrahimi, Ian Osband, Zheng Wen, Benjamin Van Roy
- abstract@open-review(Poster): We study the use of hypermodels to represent epistemic uncertainty and guide exploration. This generalizes and extends the use of ensembles to approximate Thompson sampling. The computational cost of training an ensemble grows with its size, and as such, prior work has typically been limited to ensembles with tens of elements. We show that alternative hypermodels can enjoy dramatic efficiency gains, enabling behavior that would otherwise require hundreds or thousands of elements, and even succeed in situations where ensemble methods fail to learn regardless of size. This allows more accurate approximation of Thompson sampling as well as use of more sophisticated exploration schemes. In particular, we consider an approximate form of information-directed sampling and demonstrate performance gains relative to Thompson sampling. As alternatives to ensembles, we consider linear and neural network hypermodels, also known as hypernetworks. We prove that, with neural network base models, a linear hypermodel can represent essentially any distribution over functions, and as such, hypernetworks do not extend what can be represented.

Meta Reinforcement Learning with Autonomous Inference of Subtask Dependencies

- Sungryull Sohn, Hyunjae Woo, Jongwook Choi, Honglak Lee
- abstract@open-review(Poster): We propose and address a novel few-shot RL problem, where a task is characterized by a subtask graph which describes a set of subtasks and their dependencies that are unknown to the agent. The agent needs to quickly adapt to the task over few episodes during adaptation phase to maximize the return in the test phase. Instead of directly learning a meta-policy, we develop a Meta-learner with Subtask Graph Inference (MSGI), which infers the latent parameter of the task by interacting with the environment and maximizes the return given the latent parameter. To facilitate learning, we adopt an intrinsic reward inspired by upper confidence bound (UCB) that encourages efficient exploration. Our experiment results on two grid-world domains and StarCraft II environments show that the proposed method is able to accurately infer the latent task parameter, and to adapt more efficiently than existing meta RL and hierarchical RL methods.

Explanation by Progressive Exaggeration

- Sumedha Singla, Brian Pollack, Junxiang Chen, Kayhan Batmanghelich
- abstract@open-review(Spotlight): As machine learning methods see greater adoption and implementation in high stakes applications such as medical image diagnosis, the need for model interpretability and explanation has become more critical. Classical approaches that assess feature importance (eg saliency maps) do not explain how and why a particular region of an image is relevant to the prediction. We propose a method that explains the outcome of a classification black-box by gradually exaggerating the semantic effect of a given class. Given a query input to a classifier, our method produces a progressive set of plausible variations of that query, which gradually change the posterior probability from its original class to its negation. These counter-factually generated samples preserve features unrelated to the classification decision, such that a user can employ our method as a ``tuning knob'' to traverse a data manifold while crossing the decision boundary. Our method is model agnostic and only requires the output value and gradient of the predictor with respect to its input.

BayesOpt Adversarial Attack

- Binxin Ru, Adam Cobb, Arno Blaas, Yarin Gal

- abstract@open-review(Poster): Black-box adversarial attacks require a large number of attempts before finding successful adversarial examples that are visually indistinguishable from the original input. Current approaches relying on substitute model training, gradient estimation or genetic algorithms often require an excessive number of queries. Therefore, they are not suitable for real-world systems where the maximum query number is limited due to cost. We propose a query-efficient black-box attack which uses Bayesian optimisation in combination with Bayesian model selection to optimise over the adversarial perturbation and the optimal degree of search space dimension reduction. We demonstrate empirically that our method can achieve comparable success rates with 2-5 times fewer queries compared to previous state-of-the-art black-box attacks.

Directional Message Passing for Molecular Graphs

- Johannes Gasteiger, Janek Groß, Stephan Günnemann
- abstract@open-review(Spotlight): Graph neural networks have recently achieved great successes in predicting quantum mechanical properties of molecules. These models represent a molecule as a graph using only the distance between atoms (nodes). They do not, however, consider the spatial direction from one atom to another, despite directional information playing a central role in empirical potentials for molecules, e.g. in angular potentials. To alleviate this limitation we propose directional message passing, in which we embed the messages passed between atoms instead of the atoms themselves. Each message is associated with a direction in coordinate space. These directional message embeddings are rotationally equivariant since the associated directions rotate with the molecule. We propose a message passing scheme analogous to belief propagation, which uses the directional information by transforming messages based on the angle between them. Additionally, we use spherical Bessel functions and spherical harmonics to construct theoretically well-founded, orthogonal representations that achieve better performance than the currently prevalent Gaussian radial basis representations while using fewer than 1/4 of the parameters. We leverage these innovations to construct the directional message passing neural network (DimeNet). DimeNet outperforms previous GNNs on average by 76% on MD17 and by 31% on QM9. Our implementation is available online.

Model-based reinforcement learning for biological sequence design

- Christof Angermueller, David Dohan, David Belanger, Ramya Deshpande, Kevin Murphy, Lucy Colwell
- abstract@open-review(Poster): The ability to design biological structures such as DNA or proteins would have considerable medical and industrial impact. Doing so presents a challenging black-box optimization problem characterized by the large-batch, low round setting due to the need for labor-intensive wet lab evaluations. In response, we propose using reinforcement learning (RL) based on proximal-policy optimization (PPO) for biological sequence design. RL provides a flexible framework for optimization generative sequence models to achieve specific criteria, such as diversity among the high-quality sequences discovered. We propose a model-based variant of PPO, DyNA-PPO, to improve sample efficiency, where the policy for a new round is trained offline using a simulator fit on functional measurements from prior rounds. To accommodate the growing number of observations across rounds, the simulator model is automatically selected at each round from a pool of diverse models of varying capacity. On the tasks of designing DNA transcription factor binding sites, designing antimicrobial proteins, and optimizing the energy of Ising models based on protein structure, we find that DyNA-PPO performs significantly better than existing methods in settings in which modeling is feasible, while still not performing worse in situations in which a reliable model cannot be learned.

BinaryDuo: Reducing Gradient Mismatch in Binary Activation Network by Coupling Binary Activations

- Hyungjun Kim, Kyungsu Kim, Jinseok Kim, Jae-Joon Kim
- abstract@open-review(Poster): Binary Neural Networks (BNNs) have been garnering interest thanks to their compute cost reduction and memory savings. However, BNNs suffer from performance degradation mainly due to the gradient mismatch caused by binarizing activations. Previous works tried to address the gradient mismatch problem by reducing the discrepancy between activation functions used at forward pass and its differentiable approximation used at backward pass, which is an indirect measure. In this work, we use the gradient of smoothed loss function to better estimate the gradient mismatch in quantized neural network. Analysis using the gradient mismatch estimator indicates that using higher precision for activation is more effective than modifying the differentiable approximation of activation function. Based on the observation, we propose a new training scheme for binary activation networks called BinaryDuo in which two binary activations are coupled into a ternary activation during training. Experimental results show that BinaryDuo outperforms state-of-the-art BNNs on various benchmarks with the same amount of parameters and computing cost.

Mixed-curvature Variational Autoencoders

- Ondrej Skopek, Octavian-Eugen Ganea, Gary Bécigneul
- abstract@open-review(Poster): Euclidean space has historically been the typical workhorse geometry for machine learning applications due to its power and simplicity. However, it has recently been shown that geometric spaces with constant non-zero curvature improve representations and performance on a variety of data types and downstream tasks. Consequently, generative models like Variational Autoencoders (VAEs) have been successfully generalized to elliptical and hyperbolic latent spaces. While these approaches work well on data with particular kinds of biases e.g. tree-like data for a hyperbolic VAE, there exists no generic approach unifying and leveraging all three models. We develop a Mixed-curvature Variational Autoencoder, an efficient way to train a VAE whose latent space is a product of constant curvature Riemannian manifolds, where the per-component curvature is fixed or learnable. This generalizes the Euclidean VAE to curved latent spaces and recovers it when curvatures of all latent space components go to 0.

Demystifying Inter-Class Disentanglement

- Aviv Gabbay, Yedid Hoshen
- abstract@open-review(Poster): Learning to disentangle the hidden factors of variations within a set of observations is a key task for artificial intelligence. We present a unified formulation for class and content disentanglement and use it to illustrate the limitations of current methods. We therefore introduce LORD, a novel method based on Latent Optimization for Representation Disentanglement. We find that latent optimization, along with an asymmetric noise regularization, is superior to amortized inference for achieving disentangled representations. In extensive experiments, our method is shown to achieve better disentanglement performance than both adversarial and non-adversarial methods that use the same level of supervision. We further introduce a clustering-based approach for extending our method for settings that exhibit in-class variation with promising results on the task of domain translation.

Learning from Rules Generalizing Labeled Exemplars

- Abhijeet Awasthi, Sabyasachi Ghosh, Rasna Goyal, Sunita Sarawagi
- abstract@open-review(Spotlight): In many applications labeled data is not readily available, and needs to be collected via pain-staking human supervision. We propose a rule-exemplar method for collecting human supervision to combine the efficiency of rules with the quality of instance labels. The supervision is coupled such that it is both natural for humans and synergistic for learning. We propose a training algorithm that jointly denoises rules via latent coverage variables, and trains the model through a soft implication loss over the coverage and label variables. The denoised rules and trained model are used jointly for inference. Empirical evaluation on five different tasks shows that (1) our algorithm is more accurate than several existing methods of learning from a mix of clean and noisy supervision, and (2) the coupled rule-exemplar supervision is effective in denoising rules.

Understanding the Limitations of Conditional Generative Models

- Ethan Fetaya, Joern-Henrik Jacobsen, Will Grathwohl, Richard Zemel
- abstract@open-review(Poster): Class-conditional generative models hold promise to overcome the shortcomings of their discriminative counterparts. They are a natural choice to solve discriminative tasks in a robust manner as they jointly optimize for predictive performance and accurate modeling of the input distribution. In this work, we investigate robust classification with likelihood-based generative models from a theoretical and practical perspective to investigate if they can deliver on their promises. Our analysis focuses on a spectrum of robustness properties: (1) Detection of worst-case outliers in the form of adversarial examples; (2) Detection of average-case outliers in the form of ambiguous inputs and (3) Detection of incorrectly labeled in-distribution inputs.

Our theoretical result reveals that it is impossible to guarantee detectability of adversarially-perturbed inputs even for near-optimal generative classifiers. Experimentally, we find that while we are able to train robust models for MNIST, robustness completely breaks down on CIFAR10. We relate this failure to various undesirable model properties that can be traced to the maximum likelihood training objective. Despite being a common choice in the literature, our results indicate that likelihood-based conditional generative models may are surprisingly ineffective for robust classification.

Keep Doing What Worked: Behavior Modelling Priors for Offline Reinforcement Learning

- Noah Siegel, Jost Tobias Springenberg, Felix Berkenkamp, Abbas Abdolmaleki, Michael Neunert, Thomas Lampe, Roland Hafner, Nicolas Heess, Martin Riedmiller
- abstract@open-review(Poster): Off-policy reinforcement learning algorithms promise to be applicable in settings where only a fixed data-set (batch) of environment interactions is available and no new experience can be acquired. This property makes these algorithms appealing for real world problems such as robot control. In practice, however, standard off-policy algorithms fail in the batch setting for continuous control. In this paper, we propose a simple solution to this problem. It admits the use of data generated by arbitrary behavior policies and uses a learned prior -- the advantage-weighted behavior model (ABM) -- to bias the RL policy towards actions that have previously been executed and are likely to be successful on the new task. Our method can be seen as an extension of recent work on batch-RL that enables stable learning from conflicting data-sources. We find improvements on competitive baselines in a variety of RL tasks -- including standard continuous control benchmarks and multi-task learning for simulated and real-world robots.

Empirical Bayes Transductive Meta-Learning with Synthetic Gradients

- Shell Xu Hu, Pablo Garcia Moreno, Yang Xiao, Xi Shen, Guillaume Obozinski, Neil Lawrence, Andreas Damianou
- abstract@open-review(Poster): We propose a meta-learning approach that learns from multiple tasks in a transductive setting, by leveraging the unlabeled query set in addition to the support set to generate a more powerful model for each task. To develop our framework, we revisit the empirical Bayes formulation for multi-task learning. The evidence lower bound of the marginal log-likelihood of empirical Bayes decomposes as a sum of local KL divergences between the variational posterior and the true posterior on the query set of each task. We derive a novel amortized variational inference that couples all the variational posteriors via a meta-model, which consists of a synthetic gradient network and an initialization network. Each variational posterior is derived from synthetic gradient descent to approximate the true posterior on the query set, although where we do not have access to the true gradient. Our results on the Mini-ImageNet and CIFAR-FS benchmarks for episodic few-shot classification outperform previous state-of-the-art methods. Besides, we conduct two zero-shot learning experiments to further explore the potential of the synthetic gradient.

Spike-based causal inference for weight alignment

- Jordan Guerguiev, Konrad Kording, Blake Richards
- abstract@open-review(Poster): In artificial neural networks trained with gradient descent, the weights used for processing stimuli are also used during backward passes to calculate gradients. For the real brain to approximate gradients, gradient information would have to be propagated separately, such that one set of synaptic weights is used for processing and another set is used for backward passes. This produces the so-called "weight transport problem" for biological models of learning, where the backward weights used to calculate gradients need to mirror the forward weights used to process stimuli. This weight transport problem has been considered so hard that popular proposals for biological learning assume that the backward weights are simply random, as in the feedback alignment algorithm. However, such random weights do not appear to work well for large networks. Here we show how the discontinuity introduced in a spiking system can lead to a solution to this problem. The resulting algorithm is a special case of an estimator used for causal inference in econometrics, regression discontinuity design. We show empirically that this algorithm rapidly makes the backward weights approximate the forward weights. As the backward weights become correct, this improves learning performance over feedback alignment on tasks such as Fashion-MNIST and CIFAR-10. Our results demonstrate that a simple learning rule in a spiking network can allow neurons to produce the right backward connections and thus solve the weight transport problem.

Lookahead: A Far-sighted Alternative of Magnitude-based Pruning

- Sejun Park, *Jaeho Lee*, Sangwoo Mo, Jinwoo Shin
- abstract@open-review(Poster): Magnitude-based pruning is one of the simplest methods for pruning neural networks. Despite its simplicity, magnitude-based pruning and its variants demonstrated remarkable performances for pruning modern architectures. Based on the observation that magnitude-based pruning indeed minimizes the Frobenius distortion of a linear operator corresponding to a single layer, we develop a simple pruning method, coined lookahead pruning, by extending the single layer optimization to a multi-layer optimization. Our experimental results demonstrate that the proposed method consistently outperforms magnitude-based pruning on various networks, including VGG and ResNet, particularly in the high-sparsity regime. See https://github.com/alinlab/lookahead_pruning for codes.

VariBAD: A Very Good Method for Bayes-Adaptive Deep RL via Meta-Learning

- Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, Shimon Whiteson
- abstract@open-review(Poster): Trading off exploration and exploitation in an unknown environment is key to maximising expected return during learning. A Bayes-optimal policy, which does so optimally, conditions its actions not only on the environment state but on the agent's uncertainty about the environment. Computing a Bayes-optimal policy is however intractable for all but the smallest tasks. In this paper, we introduce variational Bayes-Adaptive Deep RL (variBAD), a way to meta-learn to perform approximate inference in an unknown environment, and incorporate task uncer- tainty directly during action selection. In a grid-world domain, we illustrate how variBAD performs structured online exploration as a function of task uncertainty. We further evaluate variBAD on MuJoCo domains widely used in meta-RL and show that it achieves higher online return than existing methods.

A Generalized Training Approach for Multiagent Learning

- Paul Muller, Shayegan Omidshafiei, Mark Rowland, Karl Tuyls, Julien Perolat, Siqi Liu, Daniel Hennes, Luke Marris, Marc Lanctot, Edward Hughes, Zhe Wang, Guy Lever, Nicolas Heess, Thore Graepel, Remi Munos
- abstract@open-review(Talk): This paper investigates a population-based training regime based on game-theoretic principles called Policy-Spaced Response Oracles (PSRO). PSRO is general in the sense that it (1) encompasses well-known algorithms such as fictitious play and double oracle as special cases, and (2) in principle applies to general-sum, many-player games. Despite this, prior studies of PSRO have been focused on two-player zero-sum games, a regime where in Nash equilibria are tractably computable. In moving from two-player zero-sum games to more general settings, computation of

Nash equilibria quickly becomes infeasible. Here, we extend the theoretical underpinnings of PSRO by considering an alternative solution concept, α-Rank, which is unique (thus faces no equilibrium selection issues, unlike Nash) and applies readily to general-sum, many-player settings. We establish convergence guarantees in several games classes, and identify links between Nash equilibria and α-Rank. We demonstrate the competitive performance of α-Rank-based PSRO against an exact Nash solver-based PSRO in 2-player Kuhn and Leduc Poker. We then go beyond the reach of prior PSRO applications by considering 3- to 5-player poker games, yielding instances where α-Rank achieves faster convergence than approximate Nash solvers, thus establishing it as a favorable general games solver. We also carry out an initial empirical validation in MuJoCo soccer, illustrating the feasibility of the proposed approach in another complex domain.

Making Efficient Use of Demonstrations to Solve Hard Exploration Problems

- Caglar Gulcehre, Tom Le Paine, Bobak Shahriari, Misha Denil, Matt Hoffman, Hubert Soyer, Richard Tanburn, Steven Kapturowski, Neil Rabinowitz, Duncan Williams, Gabriel Barth-Maron, Ziyu Wang, Nando de Freitas, Worlds Team
- abstract@open-review(Poster): This paper introduces R2D3, an agent that makes efficient use of demonstrations to solve hard exploration problems in partially observable environments with highly variable initial conditions. We also introduce a suite of eight tasks that combine these three properties, and show that R2D3 can solve several of the tasks where other state of the art methods (both with and without demonstrations) fail to see even a single successful trajectory after tens of billions of steps of exploration.

Training individually fair ML models with sensitive subspace robustness

- Mikhail Yurochkin, Amanda Bower, Yuekai Sun
- abstract@open-review(Spotlight): We consider training machine learning models that are fair in the sense that their performance is invariant under certain sensitive perturbations to the inputs. For example, the performance of a resume screening system should be invariant under changes to the gender and/or ethnicity of the applicant. We formalize this notion of algorithmic fairness as a variant of individual fairness and develop a distributionally robust optimization approach to enforce it during training. We also demonstrate the effectiveness of the approach on two ML tasks that are susceptible to gender and racial biases.

Meta-learning curiosity algorithms

- Ferran Alet, *Martin F. Schneider*, Tomas Lozano-Perez, Leslie Pack Kaelbling
- abstract@open-review(Poster): We hypothesize that curiosity is a mechanism found by evolution that encourages meaningful exploration early in an agent's life in order to expose it to experiences that enable it to obtain high rewards over the course of its lifetime. We formulate the problem of generating curious behavior as one of meta-learning: an outer loop will search over a space of curiosity mechanisms that dynamically adapt the agent's reward signal, and an inner loop will perform standard reinforcement learning using the adapted reward signal. However, current meta-RL methods based on transferring neural network weights have only generalized between very similar tasks. To broaden the generalization, we instead propose to meta-learn algorithms: pieces of code similar to those designed by humans in ML papers. Our rich language of programs combines neural networks with other building blocks such as buffers, nearest-neighbor modules and custom loss functions. We demonstrate the effectiveness of the approach empirically, finding two novel curiosity algorithms that perform on par or better than human-designed published curiosity algorithms in domains as disparate as grid navigation with image inputs, acrobot, lunar lander, ant and hopper.

vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations

- Alexei Baevski, Steffen Schneider, Michael Auli
- abstract@open-review(Poster): We propose vq-wav2vec to learn discrete representations of audio segments through a wav2vec-style self-supervised context prediction task. The algorithm uses either a gumbel softmax or online k-means clustering to quantize the dense representations. Discretization enables the direct application of algorithms from the NLP community which require discrete inputs. Experiments show that BERT pre-training achieves a new state of the art on TIMIT phoneme classification and WSJ speech recognition.

Infinite-horizon Off-Policy Policy Evaluation with Multiple Behavior Policies

- Xinyun Chen, Lu Wang, Yizhe Hang, Heng Ge, Hongyuan Zha
- abstract@open-review(Poster): We consider off-policy policy evaluation when the trajectory data are generated by multiple behavior policies. Recent work has shown the key role played by the state or state-action stationary distribution corrections in the infinite horizon context for off-policy policy evaluation. We propose estimated mixture policy (EMP), a novel class of partially policy-agnostic methods to accurately estimate those quantities. With careful analysis, we show that EMP gives rise to estimates with reduced variance for estimating the state stationary distribution correction while it also offers a useful induction bias for estimating the state-action stationary distribution correction. In extensive experiments with both continuous and discrete environments, we demonstrate that our algorithm offers significantly improved accuracy compared to the state-of-the-art methods.

MMA Training: Direct Input Space Margin Maximization through Adversarial Training

- Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, Ruitong Huang
- abstract@open-review(Poster): We study adversarial robustness of neural networks from a margin maximization perspective, where margins are defined as the distances from inputs to a classifier's decision boundary. Our study shows that maximizing margins can be achieved by minimizing the adversarial loss on the decision boundary at the "shortest successful perturbation", demonstrating a close connection between adversarial losses and the margins. We propose Max-Margin Adversarial (MMA) training to directly maximize the margins to achieve adversarial robustness. Instead of adversarial training with a fixed ϵ, MMA offers an improvement by enabling adaptive selection of the "correct" ϵ as the margin individually for each datapoint. In addition, we rigorously analyze adversarial training with the perspective of margin maximization, and provide an alternative interpretation for adversarial training, maximizing either a lower or an upper bound of the margins. Our experiments empirically confirm our theory and demonstrate MMA training's efficacy on the MNIST and CIFAR10 datasets w.r.t. ℓ_∞ and ℓ_2 robustness.

Building Deep Equivariant Capsule Networks

- Sai Raam Venkataraman, S. Balasubramanian, R. Raghunatha Sarma
- abstract@open-review(Talk): Capsule networks are constrained by the parameter-expensive nature of their layers, and the general lack of provable equivariance guarantees. We present a variation of capsule networks that aims to remedy this. We identify that learning all pair-wise part-whole relationships between capsules of successive layers is inefficient. Further, we also realise that the choice of prediction networks and the routing mechanism are both key to equivariance. Based on these, we propose an alternative framework for capsule networks that learns to projectively encode the manifold of pose-variations, termed the space-of-variation (SOV), for every capsule-type of each layer. This is done using a trainable, equivariant function defined over a grid of group-transformations. Thus, the prediction-phase of routing involves projection into the SOV of a deeper capsule using the corresponding function. As a specific instantiation of this idea, and also in order to reap the benefits of increased parameter-sharing, we use type-homogeneous group-equivariant convolutions of shallower capsules in this phase. We also introduce an equivariant routing mechanism based on degree-centrality. We show that this particular instance of our general model is equivariant, and hence preserves the compositional representation of an input

under transformations. We conduct several experiments on standard object-classification datasets that showcase the increased transformation-robustness, as well as general performance, of our model to several capsule baselines.

Incorporating BERT into Neural Machine Translation

- Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, Tieyan Liu
- abstract@open-review(Poster): The recently proposed BERT (Devlin et al., 2019) has shown great power on a variety of natural language understanding tasks, such as text classification, reading comprehension, etc. However, how to effectively apply BERT to neural machine translation (NMT) lacks enough exploration. While BERT is more commonly used as fine-tuning instead of contextual embedding for downstream language understanding tasks, in NMT, our preliminary exploration of using BERT as contextual embedding is better than using for fine-tuning. This motivates us to think how to better leverage BERT for NMT along this direction. We propose a new algorithm named BERT-fused model, in which we first use BERT to extract representations for an input sequence, and then the representations are fused with each layer of the encoder and decoder of the NMT model through attention mechanisms. We conduct experiments on supervised (including sentence-level and document-level translations), semi-supervised and unsupervised machine translation, and achieve state-of-the-art results on seven benchmark datasets. Our code is available at https://github.com/bert-nmt/bert-nmt

What Can Neural Networks Reason About?

- Keyulu Xu, Jingling Li, Mozhi Zhang, Simon S. Du, Ken-ichi Kawarabayashi, Stefanie Jegelka
- abstract@open-review(Spotlight): Neural networks have succeeded in many reasoning tasks. Empirically, these tasks require specialized network structures, e.g., Graph Neural Networks (GNNs) perform well on many such tasks, but less structured networks fail. Theoretically, there is limited understanding of why and when a network structure generalizes better than others, although they have equal expressive power. In this paper, we develop a framework to characterize which reasoning tasks a network can learn well, by studying how well its computation structure aligns with the algorithmic structure of the relevant reasoning process. We formally define this algorithmic alignment and derive a sample complexity bound that decreases with better alignment. This framework offers an explanation for the empirical success of popular reasoning models, and suggests their limitations. As an example, we unify seemingly different reasoning tasks, such as intuitive physics, visual question answering, and shortest paths, via the lens of a powerful algorithmic paradigm, dynamic programming (DP). We show that GNNs align with DP and thus are expected to solve these tasks. On several reasoning tasks, our theory is supported by empirical results.

Structured Object-Aware Physics Prediction for Video Modeling and Planning

- Jannik Kossen, Karl Stelzner, Marcel Hussing, Claas Voelcker, Kristian Kersting
- abstract@open-review(Poster): When humans observe a physical system, they can easily locate components, understand their interactions, and anticipate future behavior, even in settings with complicated and previously unseen interactions. For computers, however, learning such models from videos in an unsupervised fashion is an unsolved research problem. In this paper, we present STOVE, a novel state-space model for videos, which explicitly reasons about objects and their positions, velocities, and interactions. It is constructed by combining an image model and a dynamics model in compositional manner and improves on previous work by reusing the dynamics model for inference, accelerating and regularizing training. STOVE predicts videos with convincing physical behavior over hundreds of timesteps, outperforms previous unsupervised models, and even approaches the performance of supervised baselines. We further demonstrate the strength of our model as a simulator for sample efficient model-based control, in a task with heavily interacting objects.

Learning-Augmented Data Stream Algorithms

- Tanqiu Jiang, Yi Li, Honghao Lin, Yisong Ruan, David P. Woodruff
- abstract@open-review(Poster): The data stream model is a fundamental model for processing massive data sets with limited memory and fast processing time. Recently Hsu et al. (2019) incorporated machine learning techniques into the data stream model in order to learn relevant patterns in the input data. Such techniques were encapsulated by training an oracle to predict item frequencies in the streaming model. In this paper we explore the full power of such an oracle, showing that it can be applied to a wide array of problems in data streams, sometimes resulting in the first optimal bounds for such problems. Namely, we apply the oracle to counting distinct elements on the difference of streams, estimating frequency moments, estimating cascaded aggregates, and estimating moments of geometric data streams. For the distinct elements problem, we obtain the first memory-optimal algorithms. For estimating the p-th frequency moment for $0 < p < 2$ we obtain the first algorithms with optimal update time. For estimating the p-the frequency moment for $p > 2$ we obtain a quadratic saving in memory. We empirically validate our results, demonstrating also our improvements in practice.

word2ket: Space-efficient Word Embeddings inspired by Quantum Entanglement

- Aliakbar Panahi, Seyran Saeedi, Tom Arodz
- abstract@open-review(Spotlight): Deep learning natural language processing models often use vector word embeddings, such as word2vec or GloVe, to represent words. A discrete sequence of words can be much more easily integrated with downstream neural layers if it is represented as a sequence of continuous vectors. Also, semantic relationships between words, learned from a text corpus, can be encoded in the relative configurations of the embedding vectors. However, storing and accessing embedding vectors for all words in a dictionary requires large amount of space, and may stain systems with limited GPU memory. Here, we used approaches inspired by quantum computing to propose two related methods, word2ket and word2ketXS, for storing word embedding matrix during training and inference in a highly efficient way. Our approach achieves a hundred-fold or more reduction in the space required to store the embeddings with almost no relative drop in accuracy in practical natural language processing tasks.

Towards Hierarchical Importance Attribution: Explaining Compositional Semantics for Neural Sequence Models

- Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, Xiang Ren
- abstract@open-review(Spotlight): The impressive performance of neural networks on natural language processing tasks attributes to their ability to model complicated word and phrase compositions. To explain how the model handles semantic compositions, we study hierarchical explanation of neural network predictions. We identify non-additivity and context independent importance attributions within hierarchies as two desirable properties for highlighting word and phrase compositions. We show some prior efforts on hierarchical explanations, e.g. contextual decomposition, do not satisfy the desired properties mathematically, leading to inconsistent explanation quality in different models. In this paper, we start by proposing a formal and general way to quantify the importance of each word and phrase. Following the formulation, we propose Sampling and Contextual Decomposition (SCD) algorithm and Sampling and Occlusion (SOC) algorithm. Human and metrics evaluation on both LSTM models and BERT Transformer models on multiple datasets show that our algorithms outperform prior hierarchical explanation algorithms. Our algorithms help to visualize semantic composition captured by models, extract classification rules and improve human trust of models.

On the Relationship between Self-Attention and Convolutional Layers

- Jean-Baptiste Cordonnier, Andreas Loukas, Martin Jaggi
- abstract@open-review(Poster): Recent trends of incorporating attention mechanisms in vision have led researchers to reconsider the supremacy of convolutional layers as a primary building block. Beyond helping CNNs to handle long-range dependencies, Ramachandran et al. (2019) showed that

attention can completely replace convolution and achieve state-of-the-art performance on vision tasks. This raises the question: do learned attention layers operate similarly to convolutional layers? This work provides evidence that attention layers can perform convolution and, indeed, they often learn to do so in practice. Specifically, we prove that a multi-head self-attention layer with sufficient number of heads is at least as expressive as any convolutional layer. Our numerical experiments then show that self-attention layers attend to pixel-grid patterns similarly to CNN layers, corroborating our analysis. Our code is publicly available.

SpikeGrad: An ANN-equivalent Computation Model for Implementing Backpropagation with Spikes

- Johannes C. Thiele, Olivier Bichler, Antoine Dupret
- abstract@open-review(Poster): Event-based neuromorphic systems promise to reduce the energy consumption of deep neural networks by replacing expensive floating point operations on dense matrices by low energy, sparse operations on spike events. While these systems can be trained increasingly well using approximations of the backpropagation algorithm, this usually requires high precision errors and is therefore incompatible with the typical communication infrastructure of neuromorphic circuits. In this work, we analyze how the gradient can be discretized into spike events when training a spiking neural network. To accelerate our simulation, we show that using a special implementation of the integrate-and-fire neuron allows us to describe the accumulated activations and errors of the spiking neural network in terms of an equivalent artificial neural network, allowing us to largely speed up training compared to an explicit simulation of all spike events. This way we are able to demonstrate that even for deep networks, the gradients can be discretized sufficiently well with spikes if the gradient is properly rescaled. This form of spike-based backpropagation enables us to achieve equivalent or better accuracies on the MNIST and CIFAR10 datasets than comparable state-of-the-art spiking neural networks trained with full precision gradients. The algorithm, which we call SpikeGrad, is based on only accumulation and comparison operations and can naturally exploit sparsity in the gradient computation, which makes it an interesting choice for a spiking neuromorphic systems with on-chip learning capacities.

Gradient ℓ_1 Regularization for Quantization Robustness

- Milad Alizadeh, Arash Behboodi, Mart van Baalen, Christos Louizos, Tijmen Blankevoort, Max Welling
- abstract@open-review(Poster): We analyze the effect of quantizing weights and activations of neural networks on their loss and derive a simple regularization scheme that improves robustness against post-training quantization. By training quantization-ready networks, our approach enables storing a single set of weights that can be quantized on-demand to different bit-widths as energy and memory requirements of the application change. Unlike quantization-aware training using the straight-through estimator that only targets a specific bit-width and requires access to training data and pipeline, our regularization-based method paves the way for ``on the fly'' post-training quantization to various bit-widths. We show that by modeling quantization as a ℓ_∞-bounded perturbation, the first-order term in the loss expansion can be regularized using the ℓ_1-norm of gradients. We experimentally validate our method on different vision architectures on CIFAR-10 and ImageNet datasets and show that the regularization of a neural network using our method improves robustness against quantization noise.

Spectral Embedding of Regularized Block Models

- Nathan De Lara, Thomas Bonald
- abstract@open-review(Spotlight): Spectral embedding is a popular technique for the representation of graph data. Several regularization techniques have been proposed to improve the quality of the embedding with respect to downstream tasks like clustering. In this paper, we explain on a simple block model the impact of the complete graph regularization, whereby a constant is added to all entries of the adjacency matrix. Specifically, we show that the regularization forces the spectral embedding to focus on the largest blocks, making the representation less sensitive to noise or outliers. We illustrate these results on both on both synthetic and real data, showing how regularization improves standard clustering scores.

Toward Evaluating Robustness of Deep Reinforcement Learning with Continuous Control

- Tsui-Wei Weng, Krishnamurthy (Dj) Dvijotham, *Jonathan Uesato*, Kai Xiao, *Sven Gowal*, Robert Stanforth*, Pushmeet Kohli
- abstract@open-review(Poster): Deep reinforcement learning has achieved great success in many previously difficult reinforcement learning tasks, yet recent studies show that deep RL agents are also unavoidably susceptible to adversarial perturbations, similar to deep neural networks in classification tasks. Prior works mostly focus on model-free adversarial attacks and agents with discrete actions. In this work, we study the problem of continuous control agents in deep RL with adversarial attacks and propose the first two-step algorithm based on learned model dynamics. Extensive experiments on various MuJoCo domains (Cartpole, Fish, Walker, Humanoid) demonstrate that our proposed framework is much more effective and efficient than model-free based attacks baselines in degrading agent performance as well as driving agents to unsafe states.

Decentralized Deep Learning with Arbitrary Communication Compression

- Anastasia Koloskova, *Tao Lin*, Sebastian U Stich, Martin Jaggi
- abstract@open-review(Poster): Decentralized training of deep learning models is a key element for enabling data privacy and on-device learning over networks, as well as for efficient scaling to large compute clusters. As current approaches are limited by network bandwidth, we propose the use of communication compression in the decentralized training context. We show that Choco-SGD achieves linear speedup in the number of workers for arbitrary high compression ratios on general non-convex functions, and non-IID training data. We demonstrate the practical performance of the algorithm in two key scenarios: the training of deep learning models (i) over decentralized user devices, connected by a peer-to-peer network and (ii) in a datacenter.

Training Generative Adversarial Networks from Incomplete Observations using Factorised Discriminators

- Daniel Stoller, Sebastian Ewert, Simon Dixon
- abstract@open-review(Poster): Generative adversarial networks (GANs) have shown great success in applications such as image generation and inpainting. However, they typically require large datasets, which are often not available, especially in the context of prediction tasks such as image segmentation that require labels. Therefore, methods such as the CycleGAN use more easily available unlabelled data, but do not offer a way to leverage additional labelled data for improved performance. To address this shortcoming, we show how to factorise the joint data distribution into a set of lower-dimensional distributions along with their dependencies. This allows splitting the discriminator in a GAN into multiple "sub-discriminators" that can be independently trained from incomplete observations. Their outputs can be combined to estimate the density ratio between the joint real and the generator distribution, which enables training generators as in the original GAN framework. We apply our method to image generation, image segmentation and audio source separation, and obtain improved performance over a standard GAN when additional incomplete training examples are available. For the Cityscapes segmentation task in particular, our method also improves accuracy by an absolute 14.9% over CycleGAN while using only 25 additional paired examples.

Combining Q-Learning and Search with Amortized Value Estimates

- Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Tobias Pfaff, Theophane Weber, Lars Buesing, Peter W. Battaglia
- abstract@open-review(Poster): We introduce "Search with Amortized Value Estimates" (SAVE), an approach for combining model-free Q-learning with model-based Monte-Carlo Tree Search (MCTS). In SAVE, a learned prior over state-action values is used to guide MCTS, which estimates an improved set of state-action values. The new Q-estimates are then used in combination with real experience to update the prior. This effectively amortizes the value

computation performed by MCTS, resulting in a cooperative relationship between model-free learning and model-based search. SAVE can be implemented on top of any Q-learning agent with access to a model, which we demonstrate by incorporating it into agents that perform challenging physical reasoning tasks and Atari. SAVE consistently achieves higher rewards with fewer training steps, and---in contrast to typical model-based search approaches---yields strong performance with very small search budgets. By combining real experience with information computed during search, SAVE demonstrates that it is possible to improve on both the performance of model-free learning and the computational cost of planning.

Infinite-Horizon Differentiable Model Predictive Control

- Sebastian East, Marco Gallieri, Jonathan Masci, Jan Koutnik, Mark Cannon
- abstract@open-review(Poster): This paper proposes a differentiable linear quadratic Model Predictive Control (MPC) framework for safe imitation learning. The infinite-horizon cost is enforced using a terminal cost function obtained from the discrete-time algebraic Riccati equation (DARE), so that the learned controller can be proven to be stabilizing in closed-loop. A central contribution is the derivation of the analytical derivative of the solution of the DARE, thereby allowing the use of differentiation-based learning methods. A further contribution is the structure of the MPC optimization problem: an augmented Lagrangian method ensures that the MPC optimization is feasible throughout training whilst enforcing hard constraints on state and input, and a pre-stabilizing controller ensures that the MPC solution and derivatives are accurate at each iteration. The learning capabilities of the framework are demonstrated in a set of numerical studies.

Projection-Based Constrained Policy Optimization

- Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, Peter J. Ramadge
- abstract@open-review(Poster): We consider the problem of learning control policies that optimize a reward function while satisfying constraints due to considerations of safety, fairness, or other costs. We propose a new algorithm - Projection-Based Constrained Policy Optimization (PCPO), an iterative method for optimizing policies in a two-step process - the first step performs an unconstrained update while the second step reconciles the constraint violation by projecting the policy back onto the constraint set. We theoretically analyze PCPO and provide a lower bound on reward improvement, as well as an upper bound on constraint violation for each policy update. We further characterize the convergence of PCPO with projection based on two different metrics - L2 norm and Kullback-Leibler divergence. Our empirical results over several control tasks demonstrate that our algorithm achieves superior performance, averaging more than 3.5 times less constraint violation and around 15% higher reward compared to state-of-the-art methods.

Maximum Likelihood Constraint Inference for Inverse Reinforcement Learning

- Dexter R.R. Scobee, S. Shankar Sastry
- abstract@open-review(Spotlight): While most approaches to the problem of Inverse Reinforcement Learning (IRL) focus on estimating a reward function that best explains an expert agent's policy or demonstrated behavior on a control task, it is often the case that such behavior is more succinctly represented by a simple reward combined with a set of hard constraints. In this setting, the agent is attempting to maximize cumulative rewards subject to these given constraints on their behavior. We reformulate the problem of IRL on Markov Decision Processes (MDPs) such that, given a nominal model of the environment and a nominal reward function, we seek to estimate state, action, and feature constraints in the environment that motivate an agent's behavior. Our approach is based on the Maximum Entropy IRL framework, which allows us to reason about the likelihood of an expert agent's demonstrations given our knowledge of an MDP. Using our method, we can infer which constraints can be added to the MDP to most increase the likelihood of observing these demonstrations. We present an algorithm which iteratively infers the Maximum Likelihood Constraint to best explain observed behavior, and we evaluate its efficacy using both simulated behavior and recorded data of humans navigating around an obstacle.

You Only Train Once: Loss-Conditional Training of Deep Networks

- Alexey Dosovitskiy, Josip Djolonga
- abstract@open-review(Poster): In many machine learning problems, loss functions are weighted sums of several terms. A typical approach to dealing with these is to train multiple separate models with different selections of weights and then either choose the best one according to some criterion or keep multiple models if it is desirable to maintain a diverse set of solutions. This is inefficient both at training and at inference time. We propose a method that allows replacing multiple models trained on one loss function each by a single model trained on a distribution of losses. At test time a model trained this way can be conditioned to generate outputs corresponding to any loss from the training distribution of losses. We demonstrate this approach on three tasks with parametrized losses: beta-VAE, learned image compression, and fast style transfer.

Meta-Learning Acquisition Functions for Transfer Learning in Bayesian Optimization

- Michael Volpp, Lukas P. Fröhlich, Kirsten Fischer, Andreas Doerr, Stefan Falkner, Frank Hutter, Christian Daniel
- abstract@open-review(Spotlight): Transferring knowledge across tasks to improve data-efficiency is one of the open key challenges in the field of global black-box optimization. Readily available algorithms are typically designed to be universal optimizers and, therefore, often suboptimal for specific tasks. We propose a novel transfer learning method to obtain customized optimizers within the well-established framework of Bayesian optimization, allowing our algorithm to utilize the proven generalization capabilities of Gaussian processes. Using reinforcement learning to meta-train an acquisition function (AF) on a set of related tasks, the proposed method learns to extract implicit structural information and to exploit it for improved data-efficiency. We present experiments on a simulation-to-real transfer task as well as on several synthetic functions and on two hyperparameter search problems. The results show that our algorithm (1) automatically identifies structural properties of objective functions from available source tasks or simulations, (2) performs favourably in settings with both scarse and abundant source data, and (3) falls back to the performance level of general AFs if no particular structure is present.

GraphSAINT: Graph Sampling Based Inductive Learning Method

- Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, Viktor Prasanna
- abstract@open-review(Poster): Graph Convolutional Networks (GCNs) are powerful models for learning representations of attributed graphs. To scale GCNs to large graphs, state-of-the-art methods use various layer sampling techniques to alleviate the "neighbor explosion" problem during minibatch training. We propose GraphSAINT, a graph sampling based inductive learning method that improves training efficiency and accuracy in a fundamentally different way. By changing perspective, GraphSAINT constructs minibatches by sampling the training graph, rather than the nodes or edges across GCN layers. Each iteration, a complete GCN is built from the properly sampled subgraph. Thus, we ensure fixed number of well-connected nodes in all layers. We further propose normalization technique to eliminate bias, and sampling algorithms for variance reduction. Importantly, we can decouple the sampling from the forward and backward propagation, and extend GraphSAINT with many architecture variants (e.g., graph attention, jumping connection). GraphSAINT demonstrates superior performance in both accuracy and training time on five large graphs, and achieves new state-of-the-art F1 scores for PPI (0.995) and Reddit (0.970).

Efficient Probabilistic Logic Reasoning with Graph Neural Networks

- Yuyu Zhang, Xinshi Chen, Yuan Yang, Arun Ramamurthy, Bo Li, Yuan Qi, Le Song

- abstract@open-review(Poster): Markov Logic Networks (MLNs), which elegantly combine logic rules and probabilistic graphical models, can be used to address many knowledge graph problems. However, inference in MLN is computationally intensive, making the industrial-scale application of MLN very difficult. In recent years, graph neural networks (GNNs) have emerged as efficient and effective tools for large-scale graph problems. Nevertheless, GNNs do not explicitly incorporate prior logic rules into the models, and may require many labeled examples for a target task. In this paper, we explore the combination of MLNs and GNNs, and use graph neural networks for variational inference in MLN. We propose a GNN variant, named ExpressGNN, which strikes a nice balance between the representation power and the simplicity of the model. Our extensive experiments on several benchmark datasets demonstrate that ExpressGNN leads to effective and efficient probabilistic logic reasoning.

Low-dimensional statistical manifold embedding of directed graphs

- Thorben Funke, Tian Guo, Alen Lancic, Nino Antulov-Fantulin
- abstract@open-review(Poster): We propose a novel node embedding of directed graphs to statistical manifolds, which is based on a global minimization of pairwise relative entropy and graph geodesics in a non-linear way. Each node is encoded with a probability density function over a measurable space. Furthermore, we analyze the connection of the geometrical properties of such embedding and their efficient learning procedure. Extensive experiments show that our proposed embedding is better preserving the global geodesic information of graphs, as well as outperforming existing embedding models on directed graphs in a variety of evaluation metrics, in an unsupervised setting.

RIDE: Rewarding Impact-Driven Exploration for Procedurally-Generated Environments

- Roberta Raileanu, Tim Rocktäschel
- abstract@open-review(Poster): Exploration in sparse reward environments remains one of the key challenges of model-free reinforcement learning. Instead of solely relying on extrinsic rewards provided by the environment, many state-of-the-art methods use intrinsic rewards to encourage exploration. However, we show that existing methods fall short in procedurally-generated environments where an agent is unlikely to visit a state more than once. We propose a novel type of intrinsic reward which encourages the agent to take actions that lead to significant changes in its learned state representation. We evaluate our method on multiple challenging procedurally-generated tasks in MiniGrid, as well as on tasks with high-dimensional observations used in prior work. Our experiments demonstrate that this approach is more sample efficient than existing exploration methods, particularly for procedurally-generated MiniGrid environments. Furthermore, we analyze the learned behavior as well as the intrinsic reward received by our agent. In contrast to previous approaches, our intrinsic reward does not diminish during the course of training and it rewards the agent substantially more for interacting with objects that it can control.

SPACE: Unsupervised Object-Oriented Scene Representation via Spatial Attention and Decomposition

- Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, Sungjin Ahn
- abstract@open-review(Poster): The ability to decompose complex multi-object scenes into meaningful abstractions like objects is fundamental to achieve higher-level cognition. Previous approaches for unsupervised object-oriented scene representation learning are either based on spatial-attention or scene-mixture approaches and limited in scalability which is a main obstacle towards modeling real-world scenes. In this paper, we propose a generative latent variable model, called SPACE, that provides a unified probabilistic modeling framework that combines the best of spatial-attention and scene-mixture approaches. SPACE can explicitly provide factorized object representations for foreground objects while also decomposing background segments of complex morphology. Previous models are good at either of these, but not both. SPACE also resolves the scalability problems of previous methods by incorporating parallel spatial-attention and thus is applicable to scenes with a large number of objects without performance degradations. We show through experiments on Atari and 3D-Rooms that SPACE achieves the above properties consistently in comparison to SPAIR, IODINE, and GENESIS. Results of our experiments can be found on our project website: https://sites.google.com/view/space-project-page

Cross-Lingual Ability of Multilingual BERT: An Empirical Study

- Karthikeyan K, Zihan Wang, Stephen Mayhew, Dan Roth
- abstract@open-review(Poster): Recent work has exhibited the surprising cross-lingual abilities of multilingual BERT (M-BERT) -- surprising since it is trained without any cross-lingual objective and with no aligned data. In this work, we provide a comprehensive study of the contribution of different components in M-BERT to its cross-lingual ability. We study the impact of linguistic properties of the languages, the architecture of the model, and the learning objectives. The experimental study is done in the context of three typologically different languages -- Spanish, Hindi, and Russian -- and using two conceptually different NLP tasks, textual entailment and named entity recognition. Among our key conclusions is the fact that the lexical overlap between languages plays a negligible role in the cross-lingual success, while the depth of the network is an integral part of it. All our models and implementations can be found on our project page: http://cogcomp.org/page/publication_view/900.

Reducing Transformer Depth on Demand with Structured Dropout

- Angela Fan, Edouard Grave, Armand Joulin
- abstract@open-review(Poster): Overparametrized transformer networks have obtained state of the art results in various natural language processing tasks, such as machine translation, language modeling, and question answering. These models contain hundreds of millions of parameters, necessitating a large amount of computation and making them prone to overfitting. In this work, we explore LayerDrop, a form of structured dropout, which has a regularization effect during training and allows for efficient pruning at inference time. In particular, we show that it is possible to select sub-networks of any depth from one large network without having to finetune them and with limited impact on performance. We demonstrate the effectiveness of our approach by improving the state of the art on machine translation, language modeling, summarization, question answering, and language understanding benchmarks. Moreover, we show that our approach leads to small BERT-like models of higher quality than when training from scratch or using distillation.

Neural Outlier Rejection for Self-Supervised Keypoint Learning

- Jiexiong Tang, Hanme Kim, Vitor Guizilini, Sudeep Pillai, Rares Ambrus
- abstract@open-review(Poster): Identifying salient points in images is a crucial component for visual odometry, Structure-from-Motion or SLAM algorithms. Recently, several learned keypoint methods have demonstrated compelling performance on challenging benchmarks. However, generating consistent and accurate training data for interest-point detection in natural images still remains challenging, especially for human annotators. We introduce IO-Net (i.e. InlierOutlierNet), a novel proxy task for the self-supervision of keypoint detection, description and matching. By making the sampling of inlier-outlier sets from point-pair correspondences fully differentiable within the keypoint learning framework, we show that are able to simultaneously self-supervise keypoint description and improve keypoint matching. Second, we introduce KeyPointNet, a keypoint-network architecture that is especially amenable to robust keypoint detection and description. We design the network to allow local keypoint aggregation to avoid artifacts due to spatial discretizations commonly used for this task, and we improve fine-grained keypoint descriptor performance by taking advantage of efficient sub-pixel convolutions to upsample the descriptor feature-maps to a higher operating resolution. Through extensive experiments and ablative analysis, we show that the proposed self-supervised keypoint learning method greatly improves the quality of feature matching and homography estimation on challenging benchmarks over the state-of-the-art.

B-Spline CNNs on Lie groups

- Erik J Bekkers
- abstract@open-review(Poster): Group convolutional neural networks (G-CNNs) can be used to improve classical CNNs by equipping them with the geometric structure of groups. Central in the success of G-CNNs is the lifting of feature maps to higher dimensional disentangled representations, in which data characteristics are effectively learned, geometric data-augmentations are made obsolete, and predictable behavior under geometric transformations (equivariance) is guaranteed via group theory. Currently, however, the practical implementations of G-CNNs are limited to either discrete groups (that leave the grid intact) or continuous compact groups such as rotations (that enable the use of Fourier theory). In this paper we lift these limitations and propose a modular framework for the design and implementation of G-CNNs for arbitrary Lie groups. In our approach the differential structure of Lie groups is used to expand convolution kernels in a generic basis of B-splines that is defined on the Lie algebra. This leads to a flexible framework that enables localized, atrous, and deformable convolutions in G-CNNs by means of respectively localized, sparse and non-uniform B-spline expansions. The impact and potential of our approach is studied on two benchmark datasets: cancer detection in histopathology slides (PCam dataset) in which rotation equivariance plays a key role and facial landmark localization (CelebA dataset) in which scale equivariance is important. In both cases, G-CNN architectures outperform their classical 2D counterparts and the added value of atrous and localized group convolutions is studied in detail.

Quantifying Point-Prediction Uncertainty in Neural Networks via Residual Estimation with an I/O Kernel

- Xin Qiu, Elliot Meyerson, Risto Miikkulainen
- abstract@open-review(Poster): Neural Networks (NNs) have been extensively used for a wide spectrum of real-world regression tasks, where the goal is to predict a numerical outcome such as revenue, effectiveness, or a quantitative result. In many such tasks, the point prediction is not enough: the uncertainty (i.e. risk or confidence) of that prediction must also be estimated. Standard NNs, which are most often used in such tasks, do not provide uncertainty information. Existing approaches address this issue by combining Bayesian models with NNs, but these models are hard to implement, more expensive to train, and usually do not predict as accurately as standard NNs. In this paper, a new framework (RIO) is developed that makes it possible to estimate uncertainty in any pretrained standard NN. The behavior of the NN is captured by modeling its prediction residuals with a Gaussian Process, whose kernel includes both the NN's input and its output. The framework is justified theoretically and evaluated in twelve real-world datasets, where it is found to (1) provide reliable estimates of uncertainty, (2) reduce the error of the point predictions, and (3) scale well to large datasets. Given that RIO can be applied to any standard NN without modifications to model architecture or training pipeline, it provides an important ingredient for building real-world NN applications.

EMPIR: Ensembles of Mixed Precision Deep Networks for Increased Robustness Against Adversarial Attacks

- Sanchari Sen, Balaraman Ravindran, Anand Raghunathan
- abstract@open-review(Poster): Ensuring robustness of Deep Neural Networks (DNNs) is crucial to their adoption in safety-critical applications such as self-driving cars, drones, and healthcare. Notably, DNNs are vulnerable to adversarial attacks in which small input perturbations can produce catastrophic misclassifications. In this work, we propose EMPIR, ensembles of quantized DNN models with different numerical precisions, as a new approach to increase robustness against adversarial attacks. EMPIR is based on the observation that quantized neural networks often demonstrate much higher robustness to adversarial attacks than full precision networks, but at the cost of a substantial loss in accuracy on the original (unperturbed) inputs. EMPIR overcomes this limitation to achieve the ``best of both worlds", i.e., the higher unperturbed accuracies of the full precision models combined with the higher robustness of the low precision models, by composing them in an ensemble. Further, as low precision DNN models have significantly lower computational and storage requirements than full precision models, EMPIR models only incur modest compute and memory overheads compared to a single full-precision model (<25% in our evaluations). We evaluate EMPIR across a suite of DNNs for 3 different image recognition tasks (MNIST, CIFAR-10 and ImageNet) and under 4 different adversarial attacks. Our results indicate that EMPIR boosts the average adversarial accuracies by 42.6%, 15.2% and 10.5% for the DNN models trained on the MNIST, CIFAR-10 and ImageNet datasets respectively, when compared to single full-precision models, without sacrificing accuracy on the unperturbed inputs.

Learning To Explore Using Active Neural SLAM

- Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, Ruslan Salakhutdinov
- abstract@open-review(Poster): This work presents a modular and hierarchical approach to learn policies for exploring 3D environments, called `Active Neural SLAM'. Our approach leverages the strengths of both classical and learning-based methods, by using analytical path planners with learned SLAM module, and global and local policies. The use of learning provides flexibility with respect to input modalities (in the SLAM module), leverages structural regularities of the world (in global policies), and provides robustness to errors in state estimation (in local policies). Such use of learning within each module retains its benefits, while at the same time, hierarchical decomposition and modular training allow us to sidestep the high sample complexities associated with training end-to-end policies. Our experiments in visually and physically realistic simulated 3D environments demonstrate the effectiveness of our approach over past learning and geometry-based approaches. The proposed model can also be easily transferred to the PointGoal task and was the winning entry of the CVPR 2019 Habitat PointGoal Navigation Challenge.

Understanding and Improving Information Transfer in Multi-Task Learning

- Sen Wu, Hongyang R. Zhang, Christopher Ré
- abstract@open-review(Poster): We investigate multi-task learning approaches that use a shared feature representation for all tasks. To better understand the transfer of task information, we study an architecture with a shared module for all tasks and a separate output module for each task. We study the theory of this setting on linear and ReLU-activated models. Our key observation is that whether or not tasks' data are well-aligned can significantly affect the performance of multi-task learning. We show that misalignment between task data can cause negative transfer (or hurt performance) and provide sufficient conditions for positive transfer. Inspired by the theoretical insights, we show that aligning tasks' embedding layers leads to performance gains for multi-task training and transfer learning on the GLUE benchmark and sentiment analysis tasks; for example, we obtained a 2.35% GLUE score average improvement on 5 GLUE tasks over BERT LARGE using our alignment method. We also design an SVD-based task re-weighting scheme and show that it improves the robustness of multi-task training on a multi-label image dataset.

Restricting the Flow: Information Bottlenecks for Attribution

- Karl Schulz, Leon Sixt, Federico Tombari, Tim Landgraf
- abstract@open-review(Talk): Attribution methods provide insights into the decision-making of machine learning models like artificial neural networks. For a given input sample, they assign a relevance score to each individual input variable, such as the pixels of an image. In this work, we adopt the information bottleneck concept for attribution. By adding noise to intermediate feature maps, we restrict the flow of information and can quantify (in bits) how much information image regions provide. We compare our method against ten baselines using three different metrics on VGG-16 and ResNet-50, and find that our methods outperform all baselines in five out of six settings. The method’s information-theoretic foundation provides an absolute frame of reference for attribution values (bits) and a guarantee that regions scored close to zero are not necessary for the network's decision.

A Stochastic Derivative Free Optimization Method with Momentum

- Eduard Gorbunov, Adel Bibi, Ozan Sener, El Houcine Bergou, Peter Richtarik
- abstract@open-review(Poster): We consider the problem of unconstrained minimization of a smooth objective function in $\mathbb{R}^d$ in setting where only function evaluations are possible. We propose and analyze stochastic zeroth-order method with heavy ball momentum. In particular, we propose, SMTP, a momentum version of the stochastic three-point method (STP) Bergou et al. (2019). We show new complexity results for non-convex, convex and strongly convex functions. We test our method on a collection of learning to continuous control tasks on several MuJoCo Todorov et al. (2012) environments with varying difficulty and compare against STP, other state-of-the-art derivative-free optimization algorithms and against policy gradient methods. SMTP significantly outperforms STP and all other methods that we considered in our numerical experiments. Our second contribution is SMTP with importance sampling which we call SMTP_IS. We provide convergence analysis of this method for non-convex, convex and strongly convex objectives.

Intrinsically Motivated Discovery of Diverse Patterns in Self-Organizing Systems

- Chris Reinke, Mayalen Etcheverry, Pierre-Yves Oudeyer
- abstract@open-review(Talk): In many complex dynamical systems, artificial or natural, one can observe self-organization of patterns emerging from local rules. Cellular automata, like the Game of Life (GOL), have been widely used as abstract models enabling the study of various aspects of self-organization and morphogenesis, such as the emergence of spatially localized patterns. However, findings of self-organized patterns in such models have so far relied on manual tuning of parameters and initial states, and on the human eye to identify interesting patterns. In this paper, we formulate the problem of automated discovery of diverse self-organized patterns in such high-dimensional complex dynamical systems, as well as a framework for experimentation and evaluation. Using a continuous GOL as a testbed, we show that recent intrinsically-motivated machine learning algorithms (POP-IMGEPs), initially developed for learning of inverse models in robotics, can be transposed and used in this novel application area. These algorithms combine intrinsically-motivated goal exploration and unsupervised learning of goal space representations. Goal space representations describe the interesting features of patterns for which diverse variations should be discovered. In particular, we compare various approaches to define and learn goal space representations from the perspective of discovering diverse spatially localized patterns. Moreover, we introduce an extension of a state-of-the-art POP-IMGEP algorithm which incrementally learns a goal representation using a deep auto-encoder, and the use of CPPN primitives for generating initialization parameters. We show that it is more efficient than several baselines and equally efficient as a system pre-trained on a hand-made database of patterns identified by human experts.

The Ingredients of Real World Robotic Reinforcement Learning

- Henry Zhu, Justin Yu, Abhishek Gupta, Dhruv Shah, Kristian Hartikainen, Avi Singh, Vikash Kumar, Sergey Levine
- abstract@open-review(Spotlight): The success of reinforcement learning in the real world has been limited to instrumented laboratory scenarios, often requiring arduous human supervision to enable continuous learning. In this work, we discuss the required elements of a robotic system that can continually and autonomously improve with data collected in the real world, and propose a particular instantiation of such a system. Subsequently, we investigate a number of challenges of learning without instrumentation -- including the lack of episodic resets, state estimation, and hand-engineered rewards -- and propose simple, scalable solutions to these challenges. We demonstrate the efficacy of our proposed system on dexterous robotic manipulation tasks in simulation and the real world, and also provide an insightful analysis and ablation study of the challenges associated with this learning paradigm.

Causal Discovery with Reinforcement Learning

- Shengyu Zhu, Ignavier Ng, Zhitang Chen
- abstract@open-review(Talk): Discovering causal structure among a set of variables is a fundamental problem in many empirical sciences. Traditional score-based casual discovery methods rely on various local heuristics to search for a Directed Acyclic Graph (DAG) according to a predefined score function. While these methods, e.g., greedy equivalence search, may have attractive results with infinite samples and certain model assumptions, they are less satisfactory in practice due to finite data and possible violation of assumptions. Motivated by recent advances in neural combinatorial optimization, we propose to use Reinforcement Learning (RL) to search for the DAG with the best scoring. Our encoder-decoder model takes observable data as input and generates graph adjacency matrices that are used to compute rewards. The reward incorporates both the predefined score function and two penalty terms for enforcing acyclicity. In contrast with typical RL applications where the goal is to learn a policy, we use RL as a search strategy and our final output would be the graph, among all graphs generated during training, that achieves the best reward. We conduct experiments on both synthetic and real datasets, and show that the proposed approach not only has an improved search ability but also allows for a flexible score function under the acyclicity constraint.

Scaling Autoregressive Video Models

- Dirk Weissenborn, Oscar Täckström, Jakob Uszkoreit
- abstract@open-review(Spotlight): Due to the statistical complexity of video, the high degree of inherent stochasticity, and the sheer amount of data, generating natural video remains a challenging task. State-of-the-art video generation models attempt to address these issues by combining sometimes complex, often video-specific neural network architectures, latent variable models, adversarial training and a range of other methods. Despite their often high complexity, these approaches still fall short of generating high quality video continuations outside of narrow domains and often struggle with fidelity. In contrast, we show that conceptually simple, autoregressive video generation models based on a three-dimensional self-attention mechanism achieve highly competitive results across multiple metrics on popular benchmark datasets for which they produce continuations of high fidelity and realism. Furthermore, we find that our models are capable of producing diverse and surprisingly realistic continuations on a subset of videos from Kinetics, a large scale action recognition dataset comprised of YouTube videos exhibiting phenomena such as camera movement, complex object interactions and diverse human movement. To our knowledge, this is the first promising application of video-generation models to videos of this complexity.

Compressive Transformers for Long-Range Sequence Modelling

- Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, Timothy P. Lillicrap
- abstract@open-review(Poster): We present the Compressive Transformer, an attentive sequence model which compresses past memories for long-range sequence learning. We find the Compressive Transformer obtains state-of-the-art language modelling results in the WikiText-103 and Enwik8 benchmarks, achieving 17.1 ppl and 0.97bpc respectively. We also find it can model high-frequency speech effectively and can be used as a memory mechanism for RL, demonstrated on an object matching task. To promote the domain of long-range sequence learning, we propose a new open-vocabulary language modelling benchmark derived from books, PG-19.

Differentiation of Blackbox Combinatorial Solvers

- Marin Vlastelica Pogančić, Anselm Paulus, Vit Musil, Georg Martius, Michal Rolinek
- abstract@open-review(Spotlight): Achieving fusion of deep learning with combinatorial algorithms promises transformative changes to artificial intelligence. One possible approach is to introduce combinatorial building blocks into neural networks. Such end-to-end architectures have the potential to tackle combinatorial problems on raw input data such as ensuring global consistency in multi-object tracking or route planning on maps in robotics. In this work, we present a method that implements an efficient backward pass through blackbox implementations of combinatorial solvers with linear objective functions. We provide both theoretical and experimental backing. In particular, we incorporate the Gurobi MIP solver, Blossom V algorithm, and

Dijkstra's algorithm into architectures that extract suitable features from raw inputs for the traveling salesman problem, the min-cost perfect matching problem and the shortest path problem.

Reinforced Genetic Algorithm Learning for Optimizing Computation Graphs

- Aditya Paliwal, Felix Gimeno, Vinod Nair, Yujia Li, Miles Lubin, Pushmeet Kohli, Oriol Vinyals
- abstract@open-review(Poster): We present a deep reinforcement learning approach to minimizing the execution cost of neural network computation graphs in an optimizing compiler. Unlike earlier learning-based works that require training the optimizer on the same graph to be optimized, we propose a learning approach that trains an optimizer offline and then generalizes to previously unseen graphs without further training. This allows our approach to produce high-quality execution decisions on real-world TensorFlow graphs in seconds instead of hours. We consider two optimization tasks for computation graphs: minimizing running time and peak memory usage. In comparison to an extensive set of baselines, our approach achieves significant improvements over classical and other learning-based methods on these two tasks.

Lagrangian Fluid Simulation with Continuous Convolutions

- Benjamin Ummenhofer, Lukas Prantl, Nils Thuerey, Vladlen Koltun
- abstract@open-review(Poster): We present an approach to Lagrangian fluid simulation with a new type of convolutional network. Our networks process sets of moving particles, which describe fluids in space and time. Unlike previous approaches, we do not build an explicit graph structure to connect the particles but use spatial convolutions as the main differentiable operation that relates particles to their neighbors. To this end we present a simple, novel, and effective extension of N-D convolutions to the continuous domain. We show that our network architecture can simulate different materials, generalizes to arbitrary collision geometries, and can be used for inverse problems. In addition, we demonstrate that our continuous convolutions outperform prior formulations in terms of accuracy and speed.

Harnessing the Power of Infinitely Wide Deep Nets on Small-data Tasks

- Sanjeev Arora, Simon S. Du, Zhiyuan Li, Ruslan Salakhutdinov, Ruosong Wang, Dingli Yu
- abstract@open-review(Spotlight): Recent research shows that the following two models are equivalent: (a) infinitely wide neural networks (NNs) trained under l2 loss by gradient descent with infinitesimally small learning rate (b) kernel regression with respect to so-called Neural Tangent Kernels (NTKs) (Jacot et al., 2018). An efficient algorithm to compute the NTK, as well as its convolutional counterparts, appears in Arora et al. (2019a), which allowed studying performance of infinitely wide nets on datasets like CIFAR-10. However, super-quadratic running time of kernel methods makes them best suited for small-data tasks. We report results suggesting neural tangent kernels perform strongly on low-data tasks.
- On a standard testbed of classification/regression tasks from the UCI database, NTK SVM beats the previous gold standard, Random Forests (RF), and also the corresponding finite nets.
- On CIFAR-10 with 10 – 640 training samples, Convolutional NTK consistently beats ResNet-34 by 1% - 3%.
- On VOC07 testbed for few-shot image classification tasks on ImageNet with transfer learning (Goyal et al., 2019), replacing the linear SVM currently used with a Convolutional NTK SVM consistently improves performance.
- Comparing the performance of NTK with the finite-width net it was derived from, NTK behavior starts at lower net widths than suggested by theoretical analysis(Arora et al., 2019a). NTK’s efficacy may trace to lower variance of output.

Learning to Guide Random Search

- Ozan Sener, Vladlen Koltun
- abstract@open-review(Poster): We are interested in derivative-free optimization of high-dimensional functions. The sample complexity of existing methods is high and depends on problem dimensionality, unlike the dimensionality-independent rates of first-order methods. The recent success of deep learning suggests that many datasets lie on low-dimensional manifolds that can be represented by deep nonlinear models. We therefore consider derivative-free optimization of a high-dimensional function that lies on a latent low-dimensional manifold. We develop an online learning approach that learns this manifold while performing the optimization. In other words, we jointly learn the manifold and optimize the function. Our analysis suggests that the presented method significantly reduces sample complexity. We empirically evaluate the method on continuous optimization benchmarks and high-dimensional continuous control problems. Our method achieves significantly lower sample complexity than Augmented Random Search, Bayesian optimization, covariance matrix adaptation (CMA-ES), and other derivative-free optimization algorithms.

The intriguing role of module criticality in the generalization of deep networks

- Niladri Chatterji, Behnam Neyshabur, Hanie Sedghi
- abstract@open-review(Spotlight): We study the phenomenon that some modules of deep neural networks (DNNs) are more critical than others. Meaning that rewinding their parameter values back to initialization, while keeping other modules fixed at the trained parameters, results in a large drop in the network's performance. Our analysis reveals interesting properties of the loss landscape which leads us to propose a complexity measure, called module criticality, based on the shape of the valleys that connect the initial and final values of the module parameters. We formulate how generalization relates to the module criticality, and show that this measure is able to explain the superior generalization performance of some architectures over others, whereas, earlier measures fail to do so.

Self-labelling via simultaneous clustering and representation learning

- Asano YM., Rupprecht C., Vedaldi A.
- abstract@open-review(Spotlight): Combining clustering and representation learning is one of the most promising approaches for unsupervised learning of deep neural networks. However, doing so naively leads to ill posed learning problems with degenerate solutions. In this paper, we propose a novel and principled learning formulation that addresses these issues. The method is obtained by maximizing the information between labels and input data indices. We show that this criterion extends standard cross-entropy minimization to an optimal transport problem, which we solve efficiently for millions of input images and thousands of labels using a fast variant of the Sinkhorn-Knopp algorithm. The resulting method is able to self-label visual data so as to train highly competitive image representations without manual labels. Our method achieves state of the art representation learning performance for AlexNet and ResNet-50 on SVHN, CIFAR-10, CIFAR-100 and ImageNet and yields the first self-supervised AlexNet that outperforms the supervised Pascal VOC detection baseline.

Robust anomaly detection and backdoor attack detection via differential privacy

- Min Du, Ruoxi Jia, Dawn Song
- abstract@open-review(Poster): Outlier detection and novelty detection are two important topics for anomaly detection. Suppose the majority of a dataset are drawn from a certain distribution, outlier detection and novelty detection both aim to detect data samples that do not fit the distribution. Outliers refer to data samples within this dataset, while novelties refer to new samples. In the meantime, backdoor poisoning attacks for machine learning models are achieved through injecting poisoning samples into the training dataset, which could be regarded as “outliers” that are intentionally added by attackers. Differential privacy has been proposed to avoid leaking any individual’s information, when aggregated analysis is performed on a given dataset. It is

typically achieved by adding random noise, either directly to the input dataset, or to intermediate results of the aggregation mechanism. In this paper, we demonstrate that applying differential privacy could improve the utility of outlier detection and novelty detection, with an extension to detect poisoning samples in backdoor attacks. We first present a theoretical analysis on how differential privacy helps with the detection, and then conduct extensive experiments to validate the effectiveness of differential privacy in improving outlier detection, novelty detection, and backdoor attack detection.

Deep probabilistic subsampling for task-adaptive compressed sensing

- Iris A.M. Huijben, Bastiaan S. Veeling, Ruud J.G. van Sloun
- abstract@open-review(Poster): The field of deep learning is commonly concerned with optimizing predictive models using large pre-acquired datasets of densely sampled datapoints or signals. In this work, we demonstrate that the deep learning paradigm can be extended to incorporate a subsampling scheme that is jointly optimized under a desired minimum sample rate. We present Deep Probabilistic Subsampling (DPS), a widely applicable framework for task-adaptive compressed sensing that enables end-to end optimization of an optimal subset of signal samples with a subsequent model that performs a required task. We demonstrate strong performance on reconstruction and classification tasks of a toy dataset, MNIST, and CIFAR10 under stringent subsampling rates in both the pixel and the spatial frequency domain. Due to the task-agnostic nature of the framework, DPS is directly applicable to all real-world domains that benefit from sample rate reduction.

Neural Tangents: Fast and Easy Infinite Neural Networks in Python

- Roman Novak, Lechao Xiao, Jiri Hron, Jaehoon Lee, Alexander A. Alemi, Jascha Sohl-Dickstein, Samuel S. Schoenholz
- abstract@open-review(Spotlight): Neural Tangents is a library for working with infinite-width neural networks. It provides a high-level API for specifying complex and hierarchical neural network architectures. These networks can then be trained and evaluated either at finite-width as usual or in their infinite-width limit. Infinite-width networks can be trained analytically using exact Bayesian inference or using gradient descent via the Neural Tangent Kernel. Additionally, Neural Tangents provides tools to study gradient descent training dynamics of wide but finite networks in either function space or weight space.

The entire library runs out-of-the-box on CPU, GPU, or TPU. All computations can be automatically distributed over multiple accelerators with near-linear scaling in the number of devices.

In addition to the repository below, we provide an accompanying interactive Colab notebook at https://colab.research.google.com/github/google/neural-tangents/blob/master/notebooks/neural_tangents_cookbook.ipynb

Learning Robust Representations via Multi-View Information Bottleneck

- Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, Zeynep Akata
- abstract@open-review(Poster): The information bottleneck principle provides an information-theoretic method for representation learning, by training an encoder to retain all information which is relevant for predicting the label while minimizing the amount of other, excess information in the representation. The original formulation, however, requires labeled data to identify the superfluous information. In this work, we extend this ability to the multi-view unsupervised setting, where two views of the same underlying entity are provided but the label is unknown. This enables us to identify superfluous information as that not shared by both views. A theoretical analysis leads to the definition of a new multi-view model that produces state-of-the-art results on the Sketchy dataset and label-limited versions of the MIR-Flickr dataset. We also extend our theory to the single-view setting by taking advantage of standard data augmentation techniques, empirically showing better generalization capabilities when compared to common unsupervised approaches for representation learning.

Batch-shaping for learning conditional channel gated networks

- Babak Ehteshami Bejnordi, Tijmen Blankevoort, Max Welling
- abstract@open-review(Poster): We present a method that trains large capacity neural networks with significantly improved accuracy and lower dynamic computational cost. This is achieved by gating the deep-learning architecture on a fine-grained-level. Individual convolutional maps are turned on/off conditionally on features in the network. To achieve this, we introduce a new residual block architecture that gates convolutional channels in a fine-grained manner. We also introduce a generally applicable tool batch-shaping that matches the marginal aggregate posteriors of features in a neural network to a pre-specified prior distribution. We use this novel technique to force gates to be more conditional on the data. We present results on CIFAR-10 and ImageNet datasets for image classification, and Cityscapes for semantic segmentation. Our results show that our method can slim down large architectures conditionally, such that the average computational cost on the data is on par with a smaller architecture, but with higher accuracy. In particular, on ImageNet, our ResNet50 and ResNet34 gated networks obtain 74.60% and 72.55% top-1 accuracy compared to the 69.76% accuracy of the baseline ResNet18 model, for similar complexity. We also show that the resulting networks automatically learn to use more features for difficult examples and fewer features for simple examples.

Rotation-invariant clustering of neuronal responses in primary visual cortex

- Ivan Ustyuzhaninov, Santiago A. Cadena, Emmanouil Froudarakis, Paul G. Fahey, Edgar Y. Walker, Erick Cobos, Jacob Reimer, Fabian H. Sinz, Andreas S. Tolias, Matthias Bethge, Alexander S. Ecker
- abstract@open-review(Talk): Similar to a convolutional neural network (CNN), the mammalian retina encodes visual information into several dozen nonlinear feature maps, each formed by one ganglion cell type that tiles the visual space in an approximately shift-equivariant manner. Whether such organization into distinct cell types is maintained at the level of cortical image processing is an open question. Predictive models building upon convolutional features have been shown to provide state-of-the-art performance, and have recently been extended to include rotation equivariance in order to account for the orientation selectivity of V1 neurons. However, generally no direct correspondence between CNN feature maps and groups of individual neurons emerges in these models, thus rendering it an open question whether V1 neurons form distinct functional clusters. Here we build upon the rotation-equivariant representation of a CNN-based V1 model and propose a methodology for clustering the representations of neurons in this model to find functional cell types independent of preferred orientations of the neurons. We apply this method to a dataset of 6000 neurons and visualize the preferred stimuli of the resulting clusters. Our results highlight the range of non-linear computations in mouse V1.

Inductive and Unsupervised Representation Learning on Graph Structured Objects

- Lichen Wang, Bo Zong, Qianqian Ma, Wei Cheng, Jingchao Ni, Wenchao Yu, Yanchi Liu, Dongjin Song, Haifeng Chen, Yun Fu
- abstract@open-review(Poster): Inductive and unsupervised graph learning is a critical technique for predictive or information retrieval tasks where label information is difficult to obtain. It is also challenging to make graph learning inductive and unsupervised at the same time, as learning processes guided by reconstruction error based loss functions inevitably demand graph similarity evaluation that is usually computationally intractable. In this paper, we propose a general framework SEED (Sampling, Encoding, and Embedding Distributions) for inductive and unsupervised representation learning on graph structured objects. Instead of directly dealing with the computational challenges raised by graph similarity evaluation, given an input graph, the SEED framework samples a number of subgraphs whose reconstruction errors could be efficiently evaluated, encodes the subgraph samples into a collection of subgraph vectors, and employs the embedding of the subgraph vector distribution as the output vector representation for the input graph. By theoretical

analysis, we demonstrate the close connection between SEED and graph isomorphism. Using public benchmark datasets, our empirical study suggests the proposed SEED framework is able to achieve up to 10% improvement, compared with competitive baseline methods.

U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation

- Junho Kim, Minjae Kim, Hyeonwoo Kang, Kwang Hee Lee
- abstract@open-review(Poster): We propose a novel method for unsupervised image-to-image translation, which incorporates a new attention module and a new learnable normalization function in an end-to-end manner. The attention module guides our model to focus on more important regions distinguishing between source and target domains based on the attention map obtained by the auxiliary classifier. Unlike previous attention-based method which cannot handle the geometric changes between domains, our model can translate both images requiring holistic changes and images requiring large shape changes. Moreover, our new AdaLIN (Adaptive Layer-Instance Normalization) function helps our attention-guided model to flexibly control the amount of change in shape and texture by learned parameters depending on datasets. Experimental results show the superiority of the proposed method compared to the existing state-of-the-art models with a fixed network architecture and hyper-parameters.

Masked Based Unsupervised Content Transfer

- Ron Mokady, Sagie Benaim, Lior Wolf, Amit Bermano
- abstract@open-review(Poster): We consider the problem of translating, in an unsupervised manner, between two domains where one contains some additional information compared to the other. The proposed method disentangles the common and separate parts of these domains and, through the generation of a mask, focuses the attention of the underlying network to the desired augmentation alone, without wastefully reconstructing the entire target. This enables state-of-the-art quality and variety of content translation, as demonstrated through extensive quantitative and qualitative evaluation. Our method is also capable of adding the separate content of different guide images and domains as well as remove existing separate content. Furthermore, our method enables weakly-supervised semantic segmentation of the separate part of each domain, where only class labels are provided. Our code is available at https://github.com/rmokady/mbu-content-tansfer.

DropEdge: Towards Deep Graph Convolutional Networks on Node Classification

- Yu Rong, Wenbing Huang, Tingyang Xu, Junzhou Huang
- abstract@open-review(Poster): Over-fitting and over-smoothing are two main obstacles of developing deep Graph Convolutional Networks (GCNs) for node classification. In particular, over-fitting weakens the generalization ability on small dataset, while over-smoothing impedes model training by isolating output representations from the input features with the increase in network depth. This paper proposes DropEdge, a novel and flexible technique to alleviate both issues. At its core, DropEdge randomly removes a certain number of edges from the input graph at each training epoch, acting like a data augmenter and also a message passing reducer. Furthermore, we theoretically demonstrate that DropEdge either reduces the convergence speed of over-smoothing or relieves the information loss caused by it. More importantly, our DropEdge is a general skill that can be equipped with many other backbone models (e.g. GCN, ResGCN, GraphSAGE, and JKNet) for enhanced performance. Extensive experiments on several benchmarks verify that DropEdge consistently improves the performance on a variety of both shallow and deep GCNs. The effect of DropEdge on preventing over-smoothing is empirically visualized and validated as well. Codes are released on~https://github.com/DropEdge/DropEdge.

Sequential Latent Knowledge Selection for Knowledge-Grounded Dialogue

- Byeongchang Kim, Jaewoo Ahn, Gunhee Kim
- abstract@open-review(Spotlight): Knowledge-grounded dialogue is a task of generating an informative response based on both discourse context and external knowledge. As we focus on better modeling the knowledge selection in the multi-turn knowledge-grounded dialogue, we propose a sequential latent variable model as the first approach to this matter. The model named sequential knowledge transformer (SKT) can keep track of the prior and posterior distribution over knowledge; as a result, it can not only reduce the ambiguity caused from the diversity in knowledge selection of conversation but also better leverage the response information for proper choice of knowledge. Our experimental results show that the proposed model improves the knowledge selection accuracy and subsequently the performance of utterance generation. We achieve the new state-of-the-art performance on Wizard of Wikipedia (Dinan et al., 2019) as one of the most large-scale and challenging benchmarks. We further validate the effectiveness of our model over existing conversation methods in another knowledge-based dialogue Holl-E dataset (Moghe et al., 2018).

Measuring the Reliability of Reinforcement Learning Algorithms

- Stephanie C.Y. Chan, Samuel Fishman, Anoop Korattikara, John Canny, Sergio Guadarrama
- abstract@open-review(Spotlight): Lack of reliability is a well-known issue for reinforcement learning (RL) algorithms. This problem has gained increasing attention in recent years, and efforts to improve it have grown substantially. To aid RL researchers and production users with the evaluation and improvement of reliability, we propose a set of metrics that quantitatively measure different aspects of reliability. In this work, we focus on variability and risk, both during training and after learning (on a fixed policy). We designed these metrics to be general-purpose, and we also designed complementary statistical tests to enable rigorous comparisons on these metrics. In this paper, we first describe the desired properties of the metrics and their design, the aspects of reliability that they measure, and their applicability to different scenarios. We then describe the statistical tests and make additional practical recommendations for reporting results. The metrics and accompanying statistical tools have been made available as an open-source library. We apply our metrics to a set of common RL algorithms and environments, compare them, and analyze the results.

Stable Rank Normalization for Improved Generalization in Neural Networks and GANs

- Amartya Sanyal, Philip H. Torr, Puneet K. Dokania
- abstract@open-review(Spotlight): Exciting new work on generalization bounds for neural networks (NN) given by Bartlett et al. (2017); Neyshabur et al. (2018) closely depend on two parameter- dependant quantities: the Lipschitz constant upper bound and the stable rank (a softer version of rank). Even though these bounds typically have minimal practical utility, they facilitate questions on whether controlling such quantities together could improve the generalization behaviour of NNs in practice. To this end, we propose stable rank normalization (SRN), a novel, provably optimal, and computationally efficient weight-normalization scheme which minimizes the stable rank of a linear operator. Surprisingly we find that SRN, despite being non-convex, can be shown to have a unique optimal solution. We provide extensive analyses across a wide variety of NNs (DenseNet, WideResNet, ResNet, Alexnet, VGG), where applying SRN to their linear layers leads to improved classification accuracy, while simultaneously showing improvements in genealization, evaluated empirically using—(a) shattering experiments (Zhang et al., 2016); and (b) three measures of sample complexity by Bartlett et al. (2017), Neyshabur et al. (2018), & Wei & Ma. Additionally, we show that, when applied to the discriminator of GANs, it improves Inception, FID, and Neural divergence scores, while learning mappings with low empirical Lipschitz constant.

Why Not to Use Zero Imputation? Correcting Sparsity Bias in Training Neural Networks

- Joonyoung Yi, Juhyuk Lee, Kwang Joon Kim, Sung Ju Hwang, Eunho Yang

- abstract@open-review(Poster): Handling missing data is one of the most fundamental problems in machine learning. Among many approaches, the simplest and most intuitive way is zero imputation, which treats the value of a missing entry simply as zero. However, many studies have experimentally confirmed that zero imputation results in suboptimal performances in training neural networks. Yet, none of the existing work has explained what brings such performance degradations. In this paper, we introduce the variable sparsity problem (VSP), which describes a phenomenon where the output of a predictive model largely varies with respect to the rate of missingness in the given input, and show that it adversarially affects the model performance. We first theoretically analyze this phenomenon and propose a simple yet effective technique to handle missingness, which we refer to as Sparsity Normalization (SN), that directly targets and resolves the VSP. We further experimentally validate SN on diverse benchmark datasets, to show that debiasing the effect of input-level sparsity improves the performance and stabilizes the training of neural networks.

Probability Calibration for Knowledge Graph Embedding Models

- Pedro Tabacof, Luca Costabello
- abstract@open-review(Poster): Knowledge graph embedding research has overlooked the problem of probability calibration. We show popular embedding models are indeed uncalibrated. That means probability estimates associated to predicted triples are unreliable. We present a novel method to calibrate a model when ground truth negatives are not available, which is the usual case in knowledge graphs. We propose to use Platt scaling and isotonic regression alongside our method. Experiments on three datasets with ground truth negatives show our contribution leads to well calibrated models when compared to the gold standard of using negatives. We get significantly better results than the uncalibrated models from all calibration methods. We show isotonic regression offers the best the performance overall, not without trade-offs. We also show that calibrated models reach state-of-the-art accuracy without the need to define relation-specific decision thresholds.

Reformer: The Efficient Transformer

- Nikita Kitaev, Lukasz Kaiser, Anselm Levskaya
- abstract@open-review(Talk): Large Transformer models routinely achieve state-of-the-art results on a number of tasks but training these models can be prohibitively costly, especially on long sequences. We introduce two techniques to improve the efficiency of Transformers. For one, we replace dot-product attention by one that uses locality-sensitive hashing, changing its complexity from O(L^2) to O($L \log L$), where L is the length of the sequence. Furthermore, we use reversible residual layers instead of the standard residuals, which allows storing activations only once in the training process instead of N times, where N is the number of layers. The resulting model, the Reformer, performs on par with Transformer models while being much more memory-efficient and much faster on long sequences.

Target-Embedding Autoencoders for Supervised Representation Learning

- Daniel Jarrett, Mihaela van der Schaar
- abstract@open-review(Talk): Autoencoder-based learning has emerged as a staple for disciplining representations in unsupervised and semi-supervised settings. This paper analyzes a framework for improving generalization in a purely supervised setting, where the target space is high-dimensional. We motivate and formalize the general framework of target-embedding autoencoders (TEA) for supervised prediction, learning intermediate latent representations jointly optimized to be both predictable from features as well as predictive of targets---encoding the prior that variations in targets are driven by a compact set of underlying factors. As our theoretical contribution, we provide a guarantee of generalization for linear TEAs by demonstrating uniform stability, interpreting the benefit of the auxiliary reconstruction task as a form of regularization. As our empirical contribution, we extend validation of this approach beyond existing static classification applications to multivariate sequence forecasting, verifying their advantage on both linear and nonlinear recurrent architectures---thereby underscoring the further generality of this framework beyond feedforward instantiations.

Watch the Unobserved: A Simple Approach to Parallelizing Monte Carlo Tree Search

- Anji Liu, Jianshu Chen, Mingze Yu, Yu Zhai, Xuewen Zhou, Ji Liu
- abstract@open-review(Talk): Monte Carlo Tree Search (MCTS) algorithms have achieved great success on many challenging benchmarks (e.g., Computer Go). However, they generally require a large number of rollouts, making their applications costly. Furthermore, it is also extremely challenging to parallelize MCTS due to its inherent sequential nature: each rollout heavily relies on the statistics (e.g., node visitation counts) estimated from previous simulations to achieve an effective exploration-exploitation tradeoff. In spite of these difficulties, we develop an algorithm, WU-UCT, to effectively parallelize MCTS, which achieves linear speedup and exhibits only limited performance loss with an increasing number of workers. The key idea in WU-UCT is a set of statistics that we introduce to track the number of on-going yet incomplete simulation queries (named as unobserved samples). These statistics are used to modify the UCT tree policy in the selection steps in a principled manner to retain effective exploration-exploitation tradeoff when we parallelize the most time-consuming expansion and simulation steps. Experiments on a proprietary benchmark and the Atari Game benchmark demonstrate the linear speedup and the superior performance of WU-UCT comparing to existing techniques.

On the Equivalence between Positional Node Embeddings and Structural Graph Representations

- Balasubramaniam Srinivasan, Bruno Ribeiro
- abstract@open-review(Poster): This work provides the first unifying theoretical framework for node (positional) embeddings and structural graph representations, bridging methods like matrix factorization and graph neural networks. Using invariant theory, we show that relationship between structural representations and node embeddings is analogous to that of a distribution and its samples. We prove that all tasks that can be performed by node embeddings can also be performed by structural representations and vice-versa. We also show that the concept of transductive and inductive learning is unrelated to node embeddings and graph representations, clearing another source of confusion in the literature. Finally, we introduce new practical guidelines to generating and using node embeddings, which further augments standard operating procedures used today.

Disagreement-Regularized Imitation Learning

- Kiante Brantley, Wen Sun, Mikael Henaff
- abstract@open-review(Spotlight): We present a simple and effective algorithm designed to address the covariate shift problem in imitation learning. It operates by training an ensemble of policies on the expert demonstration data, and using the variance of their predictions as a cost which is minimized with RL together with a supervised behavioral cloning cost. Unlike adversarial imitation methods, it uses a fixed reward function which is easy to optimize. We prove a regret bound for the algorithm which is linear in the time horizon multiplied by a coefficient which we show to be low for certain problems in which behavioral cloning fails. We evaluate our algorithm empirically across multiple pixel-based Atari environments and continuous control tasks, and show that it matches or significantly outperforms behavioral cloning and generative adversarial imitation learning.

Neural Epitome Search for Architecture-Agnostic Network Compression

- Daquan Zhou, Xiaojie Jin, Qibin Hou, Kaixin Wang, Jianchao Yang, Jiashi Feng
- abstract@open-review(Poster): Traditional compression methods including network pruning, quantization, low rank factorization and knowledge distillation all assume that network architectures and parameters should be hardwired. In this work, we propose a new perspective on network compression, i.e., network parameters can be disentangled from the architectures. From this viewpoint, we present the Neural Epitome Search (NES), a

new neural network compression approach that learns to find compact yet expressive epitomes for weight parameters of a specified network architecture end-to-end. The complete network to compress can be generated from the learned epitome via a novel transformation method that adaptively transforms the epitomes to match shapes of the given architecture. Compared with existing compression methods, NES allows the weight tensors to be independent of the architecture design and hence can achieve a good trade-off between model compression rate and performance given a specific model size constraint. Experiments demonstrate that, on ImageNet, when taking MobileNetV2 as backbone, our approach improves the full-model baseline by 1.47% in top-1 accuracy with 25% MAdd reduction and AutoML for Model Compression (AMC) by 2.5% with nearly the same compression ratio. Moreover, taking EfficientNet-B0 as baseline, our NES yields an improvement of 1.2% but are with 10% less MAdd. In particular, our method achieves a new state-of-the-art results of 77.5% under mobile settings (<350M MAdd). Code will be made publicly available.

Hyper-SAGNN: a self-attention based graph neural network for hypergraphs

- Ruochi Zhang, Yuesong Zou, Jian Ma
- abstract@open-review(Poster): Graph representation learning for hypergraphs can be utilized to extract patterns among higher-order interactions that are critically important in many real world problems. Current approaches designed for hypergraphs, however, are unable to handle different types of hypergraphs and are typically not generic for various learning tasks. Indeed, models that can predict variable-sized heterogeneous hyperedges have not been available. Here we develop a new self-attention based graph neural network called Hyper-SAGNN applicable to homogeneous and heterogeneous hypergraphs with variable hyperedge sizes. We perform extensive evaluations on multiple datasets, including four benchmark network datasets and two single-cell Hi-C datasets in genomics. We demonstrate that Hyper-SAGNN significantly outperforms state-of-the-art methods on traditional tasks while also achieving great performance on a new task called outsider identification. We believe that Hyper-SAGNN will be useful for graph representation learning to uncover complex higher-order interactions in different applications.

A Neural Dirichlet Process Mixture Model for Task-Free Continual Learning

- Soochan Lee, Junsoo Ha, Dongsu Zhang, Gunhee Kim
- abstract@open-review(Poster): Despite the growing interest in continual learning, most of its contemporary works have been studied in a rather restricted setting where tasks are clearly distinguishable, and task boundaries are known during training. However, if our goal is to develop an algorithm that learns as humans do, this setting is far from realistic, and it is essential to develop a methodology that works in a task-free manner. Meanwhile, among several branches of continual learning, expansion-based methods have the advantage of eliminating catastrophic forgetting by allocating new resources to learn new data. In this work, we propose an expansion-based approach for task-free continual learning. Our model, named Continual Neural Dirichlet Process Mixture (CN-DPM), consists of a set of neural network experts that are in charge of a subset of the data. CN-DPM expands the number of experts in a principled way under the Bayesian nonparametric framework. With extensive experiments, we show that our model successfully performs task-free continual learning for both discriminative and generative tasks such as image classification and image generation.

On Solving Minimax Optimization Locally: A Follow-the-Ridge Approach

- Yuanhao Wang, *Guodong Zhang*, Jimmy Ba
- abstract@open-review(Poster): Many tasks in modern machine learning can be formulated as finding equilibria in sequential games. In particular, two-player zero-sum sequential games, also known as minimax optimization, have received growing interest. It is tempting to apply gradient descent to solve minimax optimization given its popularity and success in supervised learning. However, it has been noted that naive application of gradient descent fails to find some local minimax and can converge to non-local-minimax points. In this paper, we propose Follow-the-Ridge (FR), a novel algorithm that provably converges to and only converges to local minimax. We show theoretically that the algorithm addresses the notorious rotational behaviour of gradient dynamics, and is compatible with preconditioning and positive momentum. Empirically, FR solves toy minimax problems and improves the convergence of GAN training compared to the recent minimax optimization algorithms.

Distributionally Robust Neural Networks

- Shiori Sagawa, *Pang Wei Koh*, Tatsunori B. Hashimoto, Percy Liang
- abstract@open-review(Poster): Overparameterized neural networks can be highly accurate on average on an i.i.d. test set, yet consistently fail on atypical groups of the data (e.g., by learning spurious correlations that hold on average but not in such groups). Distributionally robust optimization (DRO) allows us to learn models that instead minimize the worst-case training loss over a set of pre-defined groups. However, we find that naively applying group DRO to overparameterized neural networks fails: these models can perfectly fit the training data, and any model with vanishing average training loss also already has vanishing worst-case training loss. Instead, the poor worst-case performance arises from poor generalization on some groups. By coupling group DRO models with increased regularization---stronger-than-typical L2 regularization or early stopping---we achieve substantially higher worst-group accuracies, with 10-40 percentage point improvements on a natural language inference task and two image tasks, while maintaining high average accuracies. Our results suggest that regularization is important for worst-group generalization in the overparameterized regime, even if it is not needed for average generalization. Finally, we introduce a stochastic optimization algorithm for the group DRO setting and provide convergence guarantees for the new algorithm.

Kernel of CycleGAN as a principal homogeneous space

- Nikita Moriakov, Jonas Adler, Jonas Teuwen
- abstract@open-review(Poster): Unpaired image-to-image translation has attracted significant interest due to the invention of CycleGAN, a method which utilizes a combination of adversarial and cycle consistency losses to avoid the need for paired data. It is known that the CycleGAN problem might admit multiple solutions, and our goal in this paper is to analyze the space of exact solutions and to give perturbation bounds for approximate solutions. We show theoretically that the exact solution space is invariant with respect to automorphisms of the underlying probability spaces, and, furthermore, that the group of automorphisms acts freely and transitively on the space of exact solutions. We examine the case of zero pure CycleGAN loss first in its generality, and, subsequently, expand our analysis to approximate solutions for extended CycleGAN loss where identity loss term is included. In order to demonstrate that these results are applicable, we show that under mild conditions nontrivial smooth automorphisms exist. Furthermore, we provide empirical evidence that neural networks can learn these automorphisms with unexpected and unwanted results. We conclude that finding optimal solutions to the CycleGAN loss does not necessarily lead to the envisioned result in image-to-image translation tasks and that underlying hidden symmetries can render the result useless.

Don't Use Large Mini-batches, Use Local SGD

- Tao Lin, Sebastian U. Stich, Kumar Kshitij Patel, Martin Jaggi
- abstract@open-review(Poster): Mini-batch stochastic gradient methods (SGD) are state of the art for distributed training of deep neural networks. Drastic increases in the mini-batch sizes have lead to key efficiency and scalability gains in recent years. However, progress faces a major roadblock, as models trained with large batches often do not generalize well, i.e. they do not show good accuracy on new data. As a remedy, we propose a \emph{post-local} SGD and show that it significantly improves the generalization performance compared to large-batch training on standard benchmarks while enjoying the same efficiency (time-to-accuracy) and scalability. We further provide an extensive study of the communication efficiency vs. performance trade-offs associated with a host of \emph{local SGD} variants.

Provable robustness against all adversarial l_p-perturbations for $p\geq 1$

- Francesco Croce, Matthias Hein
- abstract@open-review(Poster): In recent years several adversarial attacks and defenses have been proposed. Often seemingly robust models turn out to be non-robust when more sophisticated attacks are used. One way out of this dilemma are provable robustness guarantees. While provably robust models for specific l_p-perturbation models have been developed, we show that they do not come with any guarantee against other l_q-perturbations. We propose a new regularization scheme, MMR-Universal, for ReLU networks which enforces robustness wrt l_1- \textit{and} l_∞-perturbations and show how that leads to the first provably robust models wrt any l_p-norm for $p\geq 1$.

Model Based Reinforcement Learning for Atari

- Łukasz Kaiser, Mohammad Babaeizadeh, Piotr Miłos, Błażej Osiński, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, Afroz Mohiuddin, Ryan Sepassi, George Tucker, Henryk Michalewski
- abstract@open-review(Spotlight): Model-free reinforcement learning (RL) can be used to learn effective policies for complex tasks, such as Atari games, even from image observations. However, this typically requires very large amounts of interaction -- substantially more, in fact, than a human would need to learn the same games. How can people learn so quickly? Part of the answer may be that people can learn how the game works and predict which actions will lead to desirable outcomes. In this paper, we explore how video prediction models can similarly enable agents to solve Atari games with fewer interactions than model-free methods. We describe Simulated Policy Learning (SimPLe), a complete model-based deep RL algorithm based on video prediction models and present a comparison of several model architectures, including a novel architecture that yields the best results in our setting. Our experiments evaluate SimPLe on a range of Atari games in low data regime of 100k interactions between the agent and the environment, which corresponds to two hours of real-time play. In most games SimPLe outperforms state-of-the-art model-free algorithms, in some games by over an order of magnitude.

On Universal Equivariant Set Networks

- Nimrod Segol, Yaron Lipman
- abstract@open-review(Poster): Using deep neural networks that are either invariant or equivariant to permutations in order to learn functions on unordered sets has become prevalent. The most popular, basic models are DeepSets (Zaheer et al. 2017) and PointNet (Qi et al. 2017). While known to be universal for approximating invariant functions, DeepSets and PointNet are not known to be universal when approximating equivariant set functions. On the other hand, several recent equivariant set architectures have been proven equivariant universal (Sannai et al. 2019, Keriven and Peyre 2019), however these models either use layers that are not permutation equivariant (in the standard sense) and/or use higher order tensor variables which are less practical. There is, therefore, a gap in understanding the universality of popular equivariant set models versus theoretical ones.

In this paper we close this gap by proving that: (i) PointNet is not equivariant universal; and (ii) adding a single linear transmission layer makes PointNet universal. We call this architecture PointNetST and argue it is the simplest permutation equivariant universal model known to date. Another consequence is that DeepSets is universal, and also PointNetSeg, a popular point cloud segmentation network (used e.g., in Qi et al. 2017) is universal.

The key theoretical tool used to prove the above results is an explicit characterization of all permutation equivariant polynomial layers. Lastly, we provide numerical experiments validating the theoretical results and comparing different permutation equivariant models.

Tensor Decompositions for Temporal Knowledge Base Completion

- Timothée Lacroix, Guillaume Obozinski, Nicolas Usunier
- abstract@open-review(Poster): Most algorithms for representation learning and link prediction in relational data have been designed for static data. However, the data they are applied to usually evolves with time, such as friend graphs in social networks or user interactions with items in recommender systems. This is also the case for knowledge bases, which contain facts such as (US, has president, B. Obama, [2009-2017]) that are valid only at certain points in time. For the problem of link prediction under temporal constraints, i.e., answering queries of the form (US, has president, ?, 2012), we propose a solution inspired by the canonical decomposition of tensors of order 4. We introduce new regularization schemes and present an extension of ComplEx that achieves state-of-the-art performance. Additionally, we propose a new dataset for knowledge base completion constructed from Wikidata, larger than previous benchmarks by an order of magnitude, as a new reference for evaluating temporal and non-temporal link prediction methods.

Network Randomization: A Simple Technique for Generalization in Deep Reinforcement Learning

- Kimin Lee, Kibok Lee, Jinwoo Shin, Honglak Lee
- abstract@open-review(Poster): Deep reinforcement learning (RL) agents often fail to generalize to unseen environments (yet semantically similar to trained agents), particularly when they are trained on high-dimensional state spaces, such as images. In this paper, we propose a simple technique to improve a generalization ability of deep RL agents by introducing a randomized (convolutional) neural network that randomly perturbs input observations. It enables trained agents to adapt to new domains by learning robust features invariant across varied and randomized environments. Furthermore, we consider an inference method based on the Monte Carlo approximation to reduce the variance induced by this randomization. We demonstrate the superiority of our method across 2D CoinRun, 3D DeepMind Lab exploration and 3D robotics control tasks: it significantly outperforms various regularization and data augmentation methods for the same purpose.

Robustness Verification for Transformers

- Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, Cho-Jui Hsieh
- abstract@open-review(Poster): Robustness verification that aims to formally certify the prediction behavior of neural networks has become an important tool for understanding model behavior and obtaining safety guarantees. However, previous methods can usually only handle neural networks with relatively simple architectures. In this paper, we consider the robustness verification problem for Transformers. Transformers have complex self-attention layers that pose many challenges for verification, including cross-nonlinearity and cross-position dependency, which have not been discussed in previous works. We resolve these challenges and develop the first robustness verification algorithm for Transformers. The certified robustness bounds computed by our method are significantly tighter than those by naive Interval Bound Propagation. These bounds also shed light on interpreting Transformers as they consistently reflect the importance of different words in sentiment analysis.

Fantastic Generalization Measures and Where to Find Them

- Yiding Jiang, *Behnam Neyshabur*, Hossein Mobahi, Dilip Krishnan, Samy Bengio
- abstract@open-review(Poster): Generalization of deep networks has been intensely researched in recent years, resulting in a number of theoretical bounds and empirically motivated measures. However, most papers proposing such measures only study a small set of models, leaving open the question of whether these measures are truly useful in practice. We present the first large scale study of generalization bounds and measures in deep networks. We train over two thousand CIFAR-10 networks with systematic changes in important hyper-parameters. We attempt to uncover potential causal relationships between each measure and generalization, by using rank correlation coefficient and its modified forms. We analyze the results and show that some of the studied measures are very promising for further research.

Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks

- Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, John E. Hopcroft
- abstract@open-review(Poster): Deep learning models are vulnerable to adversarial examples crafted by applying human-imperceptible perturbations on benign inputs. However, under the black-box setting, most existing adversaries often have a poor transferability to attack other defense models. In this work, from the perspective of regarding the adversarial example generation as an optimization process, we propose two new methods to improve the transferability of adversarial examples, namely Nesterov Iterative Fast Gradient Sign Method (NI-FGSM) and Scale-Invariant attack Method (SIM). NI-FGSM aims to adapt Nesterov accelerated gradient into the iterative attacks so as to effectively look ahead and improve the transferability of adversarial examples. While SIM is based on our discovery on the scale-invariant property of deep learning models, for which we leverage to optimize the adversarial perturbations over the scale copies of the input images so as to avoid "overfitting" on the white-box model being attacked and generate more transferable adversarial examples. NI-FGSM and SIM can be naturally integrated to build a robust gradient-based attack to generate more transferable adversarial examples against the defense models. Empirical results on ImageNet dataset demonstrate that our attack methods exhibit higher transferability and achieve higher attack success rates than state-of-the-art gradient-based attacks.

Weakly Supervised Disentanglement with Guarantees

- Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, Ben Poole
- abstract@open-review(Poster): Learning disentangled representations that correspond to factors of variation in real-world data is critical to interpretable and human-controllable machine learning. Recently, concerns about the viability of learning disentangled representations in a purely unsupervised manner has spurred a shift toward the incorporation of weak supervision. However, there is currently no formalism that identifies when and how weak supervision will guarantee disentanglement. To address this issue, we provide a theoretical framework to assist in analyzing the disentanglement guarantees (or lack thereof) conferred by weak supervision when coupled with learning algorithms based on distribution matching. We empirically verify the guarantees and limitations of several weak supervision methods (restricted labeling, match-pairing, and rank-pairing), demonstrating the predictive power and usefulness of our theoretical framework.

Discrepancy Ratio: Evaluating Model Performance When Even Experts Disagree on the Truth

- Igor Lovchinsky, Alon Daks, Israel Malkin, Pouya Samangouei, Ardavan Saeedi, Yang Liu, Swami Sankaranarayanan, Tomer Gafner, Ben Sternlieb, Patrick Maher, Nathan Silberman
- abstract@open-review(Poster): In most machine learning tasks unambiguous ground truth labels can easily be acquired. However, this luxury is often not afforded to many high-stakes, real-world scenarios such as medical image interpretation, where even expert human annotators typically exhibit very high levels of disagreement with one another. While prior works have focused on overcoming noisy labels during training, the question of how to evaluate models when annotators disagree about ground truth has remained largely unexplored. To address this, we propose the discrepancy ratio: a novel, task-independent and principled framework for validating machine learning models in the presence of high label noise. Conceptually, our approach evaluates a model by comparing its predictions to those of human annotators, taking into account the degree to which annotators disagree with one another. While our approach is entirely general, we show that in the special case of binary classification, our proposed metric can be evaluated in terms of simple, closed-form expressions that depend only on aggregate statistics of the labels and not on any individual label. Finally, we demonstrate how this framework can be used effectively to validate machine learning models using two real-world tasks from medical imaging. The discrepancy ratio metric reveals what conventional metrics do not: that our models not only vastly exceed the average human performance, but even exceed the performance of the best human experts in our datasets.

Abductive Commonsense Reasoning

- Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, Yejin Choi
- abstract@open-review(Poster): Abductive reasoning is inference to the most plausible explanation. For example, if Jenny finds her house in a mess when she returns from work, and remembers that she left a window open, she can hypothesize that a thief broke into her house and caused the mess, as the most plausible explanation. While abduction has long been considered to be at the core of how people interpret and read between the lines in natural language (Hobbs et al., 1988), there has been relatively little research in support of abductive natural language inference and generation. We present the first study that investigates the viability of language-based abductive reasoning. We introduce a challenge dataset, ART, that consists of over 20k commonsense narrative contexts and 200k explanations. Based on this dataset, we conceptualize two new tasks – (i) Abductive NLI: a multiple-choice question answering task for choosing the more likely explanation, and (ii) Abductive NLG: a conditional generation task for explaining given observations in natural language. On Abductive NLI, the best model achieves 68.9% accuracy, well below human performance of 91.4%. On Abductive NLG, the current best language generators struggle even more, as they lack reasoning capabilities that are trivial for humans. Our analysis leads to new insights into the types of reasoning that deep pre-trained language models fail to perform—despite their strong performance on the related but more narrowly defined task of entailment NLI—pointing to interesting avenues for future research.

Variance Reduction With Sparse Gradients

- Melih Elibol, Lihua Lei, Michael I. Jordan
- abstract@open-review(Poster): Variance reduction methods such as SVRG and SpiderBoost use a mixture of large and small batch gradients to reduce the variance of stochastic gradients. Compared to SGD, these methods require at least double the number of operations per update to model parameters. To reduce the computational cost of these methods, we introduce a new sparsity operator: The random-top-k operator. Our operator reduces computational complexity by estimating gradient sparsity exhibited in a variety of applications by combining the top-k operator and the randomized coordinate descent operator. With this operator, large batch gradients offer an extra benefit beyond variance reduction: A reliable estimate of gradient sparsity. Theoretically, our algorithm is at least as good as the best algorithm (SpiderBoost), and further excels in performance whenever the random-top-k operator captures gradient sparsity. Empirically, our algorithm consistently outperforms SpiderBoost using various models on various tasks including image classification, natural language processing, and sparse matrix factorization. We also provide empirical evidence to support the intuition behind our algorithm via a simple gradient entropy computation, which serves to quantify gradient sparsity at every iteration.

BlockSwap: Fisher-guided Block Substitution for Network Compression on a Budget

- Jack Turner, Elliot J. Crowley, Michael O'Boyle, Amos Storkey, Gavin Gray
- abstract@open-review(Poster): The desire to map neural networks to varying-capacity devices has led to the development of a wealth of compression techniques, many of which involve replacing standard convolutional blocks in a large network with cheap alternative blocks. However, not all blocks are created equally; for a required compute budget there may exist a potent combination of many different cheap blocks, though exhaustively searching for such a combination is prohibitively expensive. In this work, we develop BlockSwap: a fast algorithm for choosing networks with interleaved block types by passing a single minibatch of training data through randomly initialised networks and gauging their Fisher potential. These networks can then be used as students and distilled with the original large network as a teacher. We demonstrate the effectiveness of the chosen networks across CIFAR-10 and ImageNet for classification, and COCO for detection, and provide a comprehensive ablation study of our approach. BlockSwap quickly explores possible block configurations using a simple architecture ranking system, yielding highly competitive networks in orders of magnitude less time than most architecture search techniques (e.g. under 5 minutes on a single GPU for CIFAR-10).

RNA Secondary Structure Prediction By Learning Unrolled Algorithms

- Xinshi Chen, Yu Li, Ramzan Umarov, Xin Gao, Le Song
- abstract@open-review(Talk): In this paper, we propose an end-to-end deep learning model, called E2Efold, for RNA secondary structure prediction which can effectively take into account the inherent constraints in the problem. The key idea of E2Efold is to directly predict the RNA base-pairing matrix, and use an unrolled algorithm for constrained programming as the template for deep architectures to enforce constraints. With comprehensive experiments on benchmark datasets, we demonstrate the superior performance of E2Efold: it predicts significantly better structures compared to previous SOTA (especially for pseudoknotted structures), while being as efficient as the fastest algorithms in terms of inference time.

Learning transport cost from subset correspondence

- Ruishan Liu, Akshay Balsubramani, James Zou
- abstract@open-review(Poster): Learning to align multiple datasets is an important problem with many applications, and it is especially useful when we need to integrate multiple experiments or correct for confounding. Optimal transport (OT) is a principled approach to align datasets, but a key challenge in applying OT is that we need to specify a cost function that accurately captures how the two datasets are related. Reliable cost functions are typically not available and practitioners often resort to using hand-crafted or Euclidean cost even if it may not be appropriate. In this work, we investigate how to learn the cost function using a small amount of side information which is often available. The side information we consider captures subset correspondence---i.e. certain subsets of points in the two data sets are known to be related. For example, we may have some images labeled as cars in both datasets; or we may have a common annotated cell type in single-cell data from two batches. We develop an end-to-end optimizer (OT-SI) that differentiates through the Sinkhorn algorithm and effectively learns the suitable cost function from side information. On systematic experiments in images, marriage-matching and single-cell RNA-seq, our method substantially outperform state-of-the-art benchmarks.

Rényi Fair Inference

- Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, Meisam Razaviyayn
- abstract@open-review(Poster): Machine learning algorithms have been increasingly deployed in critical automated decision-making systems that directly affect human lives. When these algorithms are solely trained to minimize the training/test error, they could suffer from systematic discrimination against individuals based on their sensitive attributes, such as gender or race. Recently, there has been a surge in machine learning society to develop algorithms for fair machine learning. In particular, several adversarial learning procedures have been proposed to impose fairness. Unfortunately, these algorithms either can only impose fairness up to linear dependence between the variables, or they lack computational convergence guarantees. In this paper, we use Rényi correlation as a measure of fairness of machine learning models and develop a general training framework to impose fairness. In particular, we propose a min-max formulation which balances the accuracy and fairness when solved to optimality. For the case of discrete sensitive attributes, we suggest an iterative algorithm with theoretical convergence guarantee for solving the proposed min-max problem. Our algorithm and analysis are then specialized to fair classification and fair clustering problems. To demonstrate the performance of the proposed Rényi fair inference framework in practice, we compare it with well-known existing methods on several benchmark datasets. Experiments indicate that the proposed method has favorable empirical performance against state-of-the-art approaches.

Meta Dropout: Learning to Perturb Latent Features for Generalization

- Hae Beom Lee, Taewook Nam, Eunho Yang, Sung Ju Hwang
- abstract@open-review(Poster): A machine learning model that generalizes well should obtain low errors on unseen test examples. Thus, if we know how to optimally perturb training examples to account for test examples, we may achieve better generalization performance. However, obtaining such perturbation is not possible in standard machine learning frameworks as the distribution of the test data is unknown. To tackle this challenge, we propose a novel regularization method, meta-dropout, which learns to perturb the latent features of training examples for generalization in a meta-learning framework. Specifically, we meta-learn a noise generator which outputs a multiplicative noise distribution for latent features, to obtain low errors on the test instances in an input-dependent manner. Then, the learned noise generator can perturb the training examples of unseen tasks at the meta-test time for improved generalization. We validate our method on few-shot classification datasets, whose results show that it significantly improves the generalization performance of the base model, and largely outperforms existing regularization methods such as information bottleneck, manifold mixup, and information dropout.

Adversarial AutoAugment

- Xinyu Zhang, Qiang Wang, Jian Zhang, Zhao Zhong
- abstract@open-review(Poster): Data augmentation (DA) has been widely utilized to improve generalization in training deep neural networks. Recently, human-designed data augmentation has been gradually replaced by automatically learned augmentation policy. Through finding the best policy in well-designed search space of data augmentation, AutoAugment (Cubuk et al., 2019) can significantly improve validation accuracy on image classification tasks. However, this approach is not computationally practical for large-scale problems. In this paper, we develop an adversarial method to arrive at a computationally-affordable solution called Adversarial AutoAugment, which can simultaneously optimize target related object and augmentation policy search loss. The augmentation policy network attempts to increase the training loss of a target network through generating adversarial augmentation policies, while the target network can learn more robust features from harder examples to improve the generalization. In contrast to prior work, we reuse the computation in target network training for policy evaluation, and dispense with the retraining of the target network. Compared to AutoAugment, this leads to about 12x reduction in computing cost and 11x shortening in time overhead on ImageNet. We show experimental results of our approach on CIFAR-10/CIFAR-100, ImageNet, and demonstrate significant performance improvements over state-of-the-art. On CIFAR-10, we achieve a top-1 test error of 1.36%, which is the currently best performing single model. On ImageNet, we achieve a leading performance of top-1 accuracy 79.40% on ResNet-50 and 80.00% on ResNet-50-D without extra data.

Understanding Why Neural Networks Generalize Well Through GSNR of Parameters

- Jinlong Liu, Yunzhi Bai, Guoqing Jiang, Ting Chen, Huayan Wang
- abstract@open-review(Spotlight): As deep neural networks (DNNs) achieve tremendous success across many application domains, researchers tried to explore in many aspects on why they generalize well. In this paper, we provide a novel perspective on these issues using the gradient signal to noise ratio (GSNR) of parameters during training process of DNNs. The GSNR of a parameter is simply defined as the ratio between its gradient's squared mean and variance, over the data distribution. Based on several approximations, we establish a quantitative relationship between model parameters' GSNR and the generalization gap. This relationship indicates that larger GSNR during training process leads to better generalization performance. Futher, we show that, different from that of shallow models (e.g. logistic regression, support vector machines), the gradient descent optimization dynamics of DNNs naturally produces large GSNR during training, which is probably the key to DNNs’ remarkable generalization ability.

State-only Imitation with Transition Dynamics Mismatch

- Tanmay Gangwani, Jian Peng

- abstract@open-review(Poster): Imitation Learning (IL) is a popular paradigm for training agents to achieve complicated goals by leveraging expert behavior, rather than dealing with the hardships of designing a correct reward function. With the environment modeled as a Markov Decision Process (MDP), most of the existing IL algorithms are contingent on the availability of expert demonstrations in the same MDP as the one in which a new imitator policy is to be learned. This is uncharacteristic of many real-life scenarios where discrepancies between the expert and the imitator MDPs are common, especially in the transition dynamics function. Furthermore, obtaining expert actions may be costly or infeasible, making the recent trend towards state-only IL (where expert demonstrations constitute only states or observations) ever so promising. Building on recent adversarial imitation approaches that are motivated by the idea of divergence minimization, we present a new state-only IL algorithm in this paper. It divides the overall optimization objective into two subproblems by introducing an indirection step and solves the subproblems iteratively. We show that our algorithm is particularly effective when there is a transition dynamics mismatch between the expert and imitator MDPs, while the baseline IL methods suffer from performance degradation. To analyze this, we construct several interesting MDPs by modifying the configuration parameters for the MuJoCo locomotion tasks from OpenAI Gym.

Measuring and Improving the Use of Graph Information in Graph Neural Networks

- Yifan Hou, Jian Zhang, James Cheng, Kaili Ma, Richard T. B. Ma, Hongzhi Chen, Ming-Chang Yang
- abstract@open-review(Poster): Graph neural networks (GNNs) have been widely used for representation learning on graph data. However, there is limited understanding on how much performance GNNs actually gain from graph data. This paper introduces a context-surrounding GNN framework and proposes two smoothness metrics to measure the quantity and quality of information obtained from graph data. A new, improved GNN model, called CS-GNN, is then devised to improve the use of graph information based on the smoothness values of a graph. CS-GNN is shown to achieve better performance than existing methods in different types of real graphs.

A Latent Morphology Model for Open-Vocabulary Neural Machine Translation

- Duygu Ataman, Wilker Aziz, Alexandra Birch
- abstract@open-review(Spotlight): Translation into morphologically-rich languages challenges neural machine translation (NMT) models with extremely sparse vocabularies where atomic treatment of surface forms is unrealistic. This problem is typically addressed by either pre-processing words into subword units or performing translation directly at the level of characters. The former is based on word segmentation algorithms optimized using corpus-level statistics with no regard to the translation task. The latter learns directly from translation data but requires rather deep architectures. In this paper, we propose to translate words by modeling word formation through a hierarchical latent variable model which mimics the process of morphological inflection. Our model generates words one character at a time by composing two latent representations: a continuous one, aimed at capturing the lexical semantics, and a set of (approximately) discrete features, aimed at capturing the morphosyntactic function, which are shared among different surface forms. Our model achieves better accuracy in translation into three morphologically-rich languages than conventional open-vocabulary NMT methods, while also demonstrating a better generalization capacity under low to mid-resource settings.

Universal Approximation with Certified Networks

- Maximilian Baader, Matthew Mirman, Martin Vechev
- abstract@open-review(Poster): Training neural networks to be certifiably robust is critical to ensure their safety against adversarial attacks. However, it is currently very difficult to train a neural network that is both accurate and certifiably robust. In this work we take a step towards addressing this challenge. We prove that for every continuous function f, there exists a network n such that: (i) n approximates f arbitrarily close, and (ii) simple interval bound propagation of a region B through n yields a result that is arbitrarily close to the optimal output of f on B. Our result can be seen as a Universal Approximation Theorem for interval-certified ReLU networks. To the best of our knowledge, this is the first work to prove the existence of accurate, interval-certified networks.

Explain Your Move: Understanding Agent Actions Using Specific and Relevant Feature Attribution

- Nikaash Puri, Sukriti Verma, Piyush Gupta, Dhruv Kayastha, Shripad Deshmukh, Balaji Krishnamurthy, Sameer Singh
- abstract@open-review(Poster): As deep reinforcement learning (RL) is applied to more tasks, there is a need to visualize and understand the behavior of learned agents. Saliency maps explain agent behavior by highlighting the features of the input state that are most relevant for the agent in taking an action. Existing perturbation-based approaches to compute saliency often highlight regions of the input that are not relevant to the action taken by the agent. Our proposed approach, SARFA (Specific and Relevant Feature Attribution), generates more focused saliency maps by balancing two aspects (specificity and relevance) that capture different desiderata of saliency. The first captures the impact of perturbation on the relative expected reward of the action to be explained. The second downweighs irrelevant features that alter the relative expected rewards of actions other than the action to be explained. We compare SARFA with existing approaches on agents trained to play board games (Chess and Go) and Atari games (Breakout, Pong and Space Invaders). We show through illustrative examples (Chess, Atari, Go), human studies (Chess), and automated evaluation methods (Chess) that SARFA generates saliency maps that are more interpretable for humans than existing approaches. For the code release and demo videos, see: https://nikaashpuri.github.io/sarfa-saliency/.

Learning to Balance: Bayesian Meta-Learning for Imbalanced and Out-of-distribution Tasks

- Hae Beom Lee, Hayeon Lee, Donghyun Na, Saehoon Kim, Minseop Park, Eunho Yang, Sung Ju Hwang
- abstract@open-review(Talk): While tasks could come with varying the number of instances and classes in realistic settings, the existing meta-learning approaches for few-shot classification assume that number of instances per task and class is fixed. Due to such restriction, they learn to equally utilize the meta-knowledge across all the tasks, even when the number of instances per task and class largely varies. Moreover, they do not consider distributional difference in unseen tasks, on which the meta-knowledge may have less usefulness depending on the task relatedness. To overcome these limitations, we propose a novel meta-learning model that adaptively balances the effect of the meta-learning and task-specific learning within each task. Through the learning of the balancing variables, we can decide whether to obtain a solution by relying on the meta-knowledge or task-specific learning. We formulate this objective into a Bayesian inference framework and tackle it using variational inference. We validate our Bayesian Task-Adaptive Meta-Learning (Bayesian TAML) on two realistic task- and class-imbalanced datasets, on which it significantly outperforms existing meta-learning approaches. Further ablation study confirms the effectiveness of each balancing component and the Bayesian learning framework.

Deep Symbolic Superoptimization Without Human Knowledge

- Hui Shi, Yang Zhang, Xinyun Chen, Yuandong Tian, Jishen Zhao
- abstract@open-review(Poster): Deep symbolic superoptimization refers to the task of applying deep learning methods to simplify symbolic expressions. Existing approaches either perform supervised training on human-constructed datasets that defines equivalent expression pairs, or apply reinforcement learning with human-defined equivalent trans-formation actions. In short, almost all existing methods rely on human knowledge to define equivalence, which suffers from large labeling cost and learning bias, because it is almost impossible to define and comprehensive equivalent set. We thus propose HISS, a reinforcement learning framework for symbolic super-optimization that keeps human outside the loop. HISS introduces a tree-LSTM encoder-decoder network with attention to ensure tractable learning. Our experiments show that HISS can discover more simplification rules than existing human-dependent methods, and can learn meaningful embeddings for symbolic expressions, which are indicative of equivalence.

Sample Efficient Policy Gradient Methods with Recursive Variance Reduction

- Pan Xu, Felicia Gao, Quanquan Gu
- abstract@open-review(Poster): Improving the sample efficiency in reinforcement learning has been a long-standing research problem. In this work, we aim to reduce the sample complexity of existing policy gradient methods. We propose a novel policy gradient algorithm called SRVR-PG, which only requires $O(1/\epsilon^{3/2})$\footnote{$O(\cdot)$ notation hides constant factors.} episodes to find an ϵ-approximate stationary point of the nonconcave performance function $J(\boldsymbol{\theta})$ (i.e., $\boldsymbol{\theta}$ such that $\|\nabla J(\boldsymbol{\theta})\|_2^2\leq\epsilon$). This sample complexity improves the existing result $O(1/\epsilon^{5/3})$ for stochastic variance reduced policy gradient algorithms by a factor of $O(1/\epsilon^{1/6})$. In addition, we also propose a variant of SRVR-PG with parameter exploration, which explores the initial policy parameter from a prior probability distribution. We conduct numerical experiments on classic control problems in reinforcement learning to validate the performance of our proposed algorithms.

Fast Task Inference with Variational Intrinsic Successor Features

- Steven Hansen, Will Dabney, Andre Barreto, David Warde-Farley, Tom Van de Wiele, Volodymyr Mnih
- abstract@open-review(Talk): It has been established that diverse behaviors spanning the controllable subspace of a Markov decision process can be trained by rewarding a policy for being distinguishable from other policies. However, one limitation of this formulation is the difficulty to generalize beyond the finite set of behaviors being explicitly learned, as may be needed in subsequent tasks. Successor features provide an appealing solution to this generalization problem, but require defining the reward function as linear in some grounded feature space. In this paper, we show that these two techniques can be combined, and that each method solves the other's primary limitation. To do so we introduce Variational Intrinsic Successor FeatuRes (VISR), a novel algorithm which learns controllable features that can be leveraged to provide enhanced generalization and fast task inference through the successor features framework. We empirically validate VISR on the full Atari suite, in a novel setup wherein the rewards are only exposed briefly after a long unsupervised phase. Achieving human-level performance on 12 games and beating all baselines, we believe VISR represents a step towards agents that rapidly learn from limited feedback.

Certified Defenses for Adversarial Patches

- Ping-yeh Chiang, *Renkun Ni*, Ahmed Abdelkader, Chen Zhu, Christoph Studor, Tom Goldstein
- abstract@open-review(Poster): Adversarial patch attacks are among one of the most practical threat models against real-world computer vision systems. This paper studies certified and empirical defenses against patch attacks. We begin with a set of experiments showing that most existing defenses, which work by pre-processing input images to mitigate adversarial patches, are easily broken by simple white-box adversaries. Motivated by this finding, we propose the first certified defense against patch attacks, and propose faster methods for its training. Furthermore, we experiment with different patch shapes for testing, obtaining surprisingly good robustness transfer across shapes, and present preliminary results on certified defense against sparse attacks. Our complete implementation can be found on: https://github.com/Ping-C/certifiedpatchdefense.

Contrastive Representation Distillation

- Yonglong Tian, Dilip Krishnan, Phillip Isola
- abstract@open-review(Poster): Often we wish to transfer representational knowledge from one neural network to another. Examples include distilling a large network into a smaller one, transferring knowledge from one sensory modality to a second, or ensembling a collection of models into a single estimator. Knowledge distillation, the standard approach to these problems, minimizes the KL divergence between the probabilistic outputs of a teacher and student network. We demonstrate that this objective ignores important structural knowledge of the teacher network. This motivates an alternative objective by which we train a student to capture significantly more information in the teacher's representation of the data. We formulate this objective as contrastive learning. Experiments demonstrate that our resulting new objective outperforms knowledge distillation on a variety of knowledge transfer tasks, including single model compression, ensemble distillation, and cross-modal transfer. When combined with knowledge distillation, our method sets a state of the art in many transfer tasks, sometimes even outperforming the teacher network.

A FRAMEWORK FOR ROBUSTNESS CERTIFICATION OF SMOOTHED CLASSIFIERS USING F-DIVERGENCES

- Krishnamurthy (Dj) Dvijotham, Jamie Hayes, Borja Balle, Zico Kolter, Chongli Qin, Andras Gyorgy, Kai Xiao, Sven Gowal, Pushmeet Kohli
- abstract@open-review(Poster): Formal verification techniques that compute provable guarantees on properties of machine learning models, like robustness to norm-bounded adversarial perturbations, have yielded impressive results. Although most techniques developed so far require knowledge of the architecture of the machine learning model and remain hard to scale to complex prediction pipelines, the method of randomized smoothing has been shown to overcome many of these obstacles. By requiring only black-box access to the underlying model, randomized smoothing scales to large architectures and is agnostic to the internals of the network. However, past work on randomized smoothing has focused on restricted classes of smoothing measures or perturbations (like Gaussian or discrete) and has only been able to prove robustness with respect to simple norm bounds. In this paper we introduce a general framework for proving robustness properties of smoothed machine learning models in the black-box setting. Specifically, we extend randomized smoothing procedures to handle arbitrary smoothing measures and prove robustness of the smoothed classifier by using f-divergences. Our methodology improves upon the state of the art in terms of computation time or certified robustness on several image classification tasks and an audio classification task, with respect to several classes of adversarial perturbations.

Padé Activation Units: End-to-end Learning of Flexible Activation Functions in Deep Networks

- Alejandro Molina, Patrick Schramowski, Kristian Kersting
- abstract@open-review(Poster): The performance of deep network learning strongly depends on the choice of the non-linear activation function associated with each neuron. However, deciding on the best activation is non-trivial and the choice depends on the architecture, hyper-parameters, and even on the dataset. Typically these activations are fixed by hand before training. Here, we demonstrate how to eliminate the reliance on first picking fixed activation functions by using flexible parametric rational functions instead. The resulting Padé Activation Units (PAUs) can both approximate common activation functions and also learn new ones while providing compact representations. Our empirical evidence shows that end-to-end learning deep networks with PAUs can increase the predictive performance. Moreover, PAUs pave the way to approximations with provable robustness.

Learning to Retrieve Reasoning Paths over Wikipedia Graph for Question Answering

- Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, Caiming Xiong
- abstract@open-review(Poster): Answering questions that require multi-hop reasoning at web-scale necessitates retrieving multiple evidence documents, one of which often has little lexical or semantic relationship to the question. This paper introduces a new graph-based recurrent retrieval approach that learns to retrieve reasoning paths over the Wikipedia graph to answer multi-hop open-domain questions. Our retriever model trains a recurrent neural network that learns to sequentially retrieve evidence paragraphs in the reasoning path by conditioning on the previously retrieved documents. Our reader model ranks the reasoning paths and extracts the answer span included in the best reasoning path. Experimental results show state-of-the-art results in three open-domain QA datasets, showcasing the effectiveness and robustness of our method. Notably, our method achieves significant improvement in HotpotQA, outperforming the previous best model by more than 14 points.

A Baseline for Few-Shot Image Classification

- Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, Stefano Soatto
- abstract@open-review(Poster): Fine-tuning a deep network trained with the standard cross-entropy loss is a strong baseline for few-shot learning. When fine-tuned transductively, this outperforms the current state-of-the-art on standard datasets such as Mini-ImageNet, Tiered-ImageNet, CIFAR-FS and FC-100 with the same hyper-parameters. The simplicity of this approach enables us to demonstrate the first few-shot learning results on the ImageNet-21k dataset. We find that using a large number of meta-training classes results in high few-shot accuracies even for a large number of few-shot classes. We do not advocate our approach as the solution for few-shot learning, but simply use the results to highlight limitations of current benchmarks and few-shot protocols. We perform extensive studies on benchmark datasets to propose a metric that quantifies the "hardness" of a few-shot episode. This metric can be used to report the performance of few-shot algorithms in a more systematic way.

Abstract Diagrammatic Reasoning with Multiplex Graph Networks

- Duo Wang, Mateja Jamnik, Pietro Lio
- abstract@open-review(Poster): Abstract reasoning, particularly in the visual domain, is a complex human ability, but it remains a challenging problem for artificial neural learning systems. In this work we propose MXGNet, a multilayer graph neural network for multi-panel diagrammatic reasoning tasks. MXGNet combines three powerful concepts, namely, object-level representation, graph neural networks and multiplex graphs, for solving visual reasoning tasks. MXGNet first extracts object-level representations for each element in all panels of the diagrams, and then forms a multi-layer multiplex graph capturing multiple relations between objects across different diagram panels. MXGNet summarises the multiple graphs extracted from the diagrams of the task, and uses this summarisation to pick the most probable answer from the given candidates. We have tested MXGNet on two types of diagrammatic reasoning tasks, namely Diagram Syllogisms and Raven Progressive Matrices (RPM). For an Euler Diagram Syllogism task MXGNet achieves state-of-the-art accuracy of 99.8%. For PGM and RAVEN, two comprehensive datasets for RPM reasoning, MXGNet outperforms the state-of-the-art models by a considerable margin.

Environmental drivers of systematicity and generalization in a situated agent

- Felix Hill, Andrew Lampinen, Rosalia Schneider, Stephen Clark, Matthew Botvinick, James L. McClelland, Adam Santoro
- abstract@open-review(Poster): The question of whether deep neural networks are good at generalising beyond their immediate training experience is of critical importance for learning-based approaches to AI. Here, we consider tests of out-of-sample generalisation that require an agent to respond to never-seen-before instructions by manipulating and positioning objects in a 3D Unity simulated room. We first describe a comparatively generic agent architecture that exhibits strong performance on these tests. We then identify three aspects of the training regime and environment that make a significant difference to its performance: (a) the number of object/word experiences in the training set; (b) the visual invariances afforded by the agent's perspective, or frame of reference; and (c) the variety of visual input inherent in the perceptual aspect of the agent's perception. Our findings indicate that the degree of generalisation that networks exhibit can depend critically on particulars of the environment in which a given task is instantiated. They further suggest that the propensity for neural networks to generalise in systematic ways may increase if, like human children, those networks have access to many frames of richly varying, multi-modal observations as they learn.

ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators

- Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning
- abstract@open-review(Poster): Masked language modeling (MLM) pre-training methods such as BERT corrupt the input by replacing some tokens with [MASK] and then train a model to reconstruct the original tokens. While they produce good results when transferred to downstream NLP tasks, they generally require large amounts of compute to be effective. As an alternative, we propose a more sample-efficient pre-training task called replaced token detection. Instead of masking the input, our approach corrupts it by replacing some tokens with plausible alternatives sampled from a small generator network. Then, instead of training a model that predicts the original identities of the corrupted tokens, we train a discriminative model that predicts whether each token in the corrupted input was replaced by a generator sample or not. Thorough experiments demonstrate this new pre-training task is more efficient than MLM because the task is defined over all input tokens rather than just the small subset that was masked out. As a result, the contextual representations learned by our approach substantially outperform the ones learned by BERT given the same model size, data, and compute. The gains are particularly strong for small models; for example, we train a model on one GPU for 4 days that outperforms GPT (trained using 30x more compute) on the GLUE natural language understanding benchmark. Our approach also works well at scale, where it performs comparably to RoBERTa and XLNet while using less than 1/4 of their compute and outperforms them when using the same amount of compute.

Evolutionary Population Curriculum for Scaling Multi-Agent Reinforcement Learning

- Qian Long, *Zihan Zhou*, Abhinav Gupta, Fei Fang, Yi Wu†, Xiaolong Wang†
- abstract@open-review(Poster): In multi-agent games, the complexity of the environment can grow exponentially as the number of agents increases, so it is particularly challenging to learn good policies when the agent population is large. In this paper, we introduce Evolutionary Population Curriculum (EPC), a curriculum learning paradigm that scales up Multi-Agent Reinforcement Learning (MARL) by progressively increasing the population of training agents in a stage-wise manner. Furthermore, EPC uses an evolutionary approach to fix an objective misalignment issue throughout the curriculum: agents successfully trained in an early stage with a small population are not necessarily the best candidates for adapting to later stages with scaled populations. Concretely, EPC maintains multiple sets of agents in each stage, performs mix-and-match and fine-tuning over these sets and promotes the sets of agents with the best adaptability to the next stage. We implement EPC on a popular MARL algorithm, MADDPG, and empirically show that our approach consistently outperforms baselines by a large margin as the number of agents grows exponentially. The source code and videos can be found at https://sites.google.com/view/epciclr2020.

Thinking While Moving: Deep Reinforcement Learning with Concurrent Control

- Ted Xiao, Eric Jang, Dmitry Kalashnikov, Sergey Levine, Julian Ibarz, Karol Hausman, Alexander Herzog
- abstract@open-review(Poster): We study reinforcement learning in settings where sampling an action from the policy must be done concurrently with the time evolution of the controlled system, such as when a robot must decide on the next action while still performing the previous action. Much like a person or an animal, the robot must think and move at the same time, deciding on its next action before the previous one has completed. In order to develop an algorithmic framework for such concurrent control problems, we start with a continuous-time formulation of the Bellman equations, and then discretize them in a way that is aware of system delays. We instantiate this new class of approximate dynamic programming methods via a simple architectural extension to existing value-based deep reinforcement learning algorithms. We evaluate our methods on simulated benchmark tasks and a large-scale robotic grasping task where the robot must "think while moving."

Jacobian Adversarially Regularized Networks for Robustness

- Alvin Chan, Yi Tay, Yew Soon Ong, Jie Fu
- abstract@open-review(Poster): Adversarial examples are crafted with imperceptible perturbations with the intent to fool neural networks. Against such attacks, adversarial training and its variants stand as the strongest defense to date. Previous studies have pointed out that robust models that have undergone adversarial training tend to produce more salient and interpretable Jacobian matrices than their non-robust counterparts. A natural question is whether a model trained with an objective to produce salient Jacobian can result in better robustness. This paper answers this question with affirmative empirical results. We propose Jacobian Adversarially Regularized Networks (JARN) as a method to optimize the saliency of a classifier's Jacobian by

adversarially regularizing the model's Jacobian to resemble natural training images. Image classifiers trained with JARN show improved robust accuracy compared to standard models on the MNIST, SVHN and CIFAR-10 datasets, uncovering a new angle to boost robustness without using adversarial training.

Towards Verified Robustness under Text Deletion Interventions

- Johannes Welbl, Po-Sen Huang, Robert Stanforth, Sven Gowal, Krishnamurthy (Dj) Dvijotham, Martin Szummer, Pushmeet Kohli
- abstract@open-review(Poster): Neural networks are widely used in Natural Language Processing, yet despite their empirical successes, their behaviour is brittle: they are both over-sensitive to small input changes, and under-sensitive to deletions of large fractions of input text. This paper aims to tackle under-sensitivity in the context of natural language inference by ensuring that models do not become more confident in their predictions as arbitrary subsets of words from the input text are deleted. We develop a novel technique for formal verification of this specification for models based on the popular decomposable attention mechanism by employing the efficient yet effective interval bound propagation (IBP) approach. Using this method we can efficiently prove, given a model, whether a particular sample is free from the under-sensitivity problem. We compare different training methods to address under-sensitivity, and compare metrics to measure it. In our experiments on the SNLI and MNLI datasets, we observe that IBP training leads to a significantly improved verified accuracy. On the SNLI test set, we can verify 18.4% of samples, a substantial improvement over only 2.8% using standard training.

And the Bit Goes Down: Revisiting the Quantization of Neural Networks

- Pierre Stock, Armand Joulin, Rémi Gribonval, Benjamin Graham, Hervé Jégou
- abstract@open-review(Spotlight): In this paper, we address the problem of reducing the memory footprint of convolutional network architectures. We introduce a vector quantization method that aims at preserving the quality of the reconstruction of the network outputs rather than its weights. The principle of our approach is that it minimizes the loss reconstruction error for in-domain inputs. Our method only requires a set of unlabelled data at quantization time and allows for efficient inference on CPU by using byte-aligned codebooks to store the compressed weights. We validate our approach by quantizing a high performing ResNet-50 model to a memory size of 5MB (20x compression factor) while preserving a top-1 accuracy of 76.1% on ImageNet object classification and by compressing a Mask R-CNN with a 26x factor.

RGBD-GAN: Unsupervised 3D Representation Learning From Natural Image Datasets via RGBD Image Synthesis

- Atsuhiro Noguchi, Tatsuya Harada
- abstract@open-review(Poster): Understanding three-dimensional (3D) geometries from two-dimensional (2D) images without any labeled information is promising for understanding the real world without incurring annotation cost. We herein propose a novel generative model, RGBD-GAN, which achieves unsupervised 3D representation learning from 2D images. The proposed method enables camera parameter--conditional image generation and depth image generation without any 3D annotations, such as camera poses or depth. We use an explicit 3D consistency loss for two RGBD images generated from different camera parameters, in addition to the ordinal GAN objective. The loss is simple yet effective for any type of image generator such as DCGAN and StyleGAN to be conditioned on camera parameters. Through experiments, we demonstrated that the proposed method could learn 3D representations from 2D images with various generator architectures.

Provable Benefit of Orthogonal Initialization in Optimizing Deep Linear Networks

- Wei Hu, Lechao Xiao, Jeffrey Pennington
- abstract@open-review(Poster): The selection of initial parameter values for gradient-based optimization of deep neural networks is one of the most impactful hyperparameter choices in deep learning systems, affecting both convergence times and model performance. Yet despite significant empirical and theoretical analysis, relatively little has been proved about the concrete effects of different initialization schemes. In this work, we analyze the effect of initialization in deep linear networks, and provide for the first time a rigorous proof that drawing the initial weights from the orthogonal group speeds up convergence relative to the standard Gaussian initialization with iid weights. We show that for deep networks, the width needed for efficient convergence to a global minimum with orthogonal initializations is independent of the depth, whereas the width needed for efficient convergence with Gaussian initializations scales linearly in the depth. Our results demonstrate how the benefits of a good initialization can persist throughout learning, suggesting an explanation for the recent empirical successes found by initializing very deep non-linear networks according to the principle of dynamical isometry.

Implementation Matters in Deep RL: A Case Study on PPO and TRPO

- Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, Aleksander Madry
- abstract@open-review(Talk): We study the roots of algorithmic progress in deep policy gradient algorithms through a case study on two popular algorithms: Proximal Policy Optimization (PPO) and Trust Region Policy Optimization (TRPO). Specifically, we investigate the consequences of "code-level optimizations:" algorithm augmentations found only in implementations or described as auxiliary details to the core algorithm. Seemingly of secondary importance, such optimizations turn out to have a major impact on agent behavior. Our results show that they (a) are responsible for most of PPO's gain in cumulative reward over TRPO, and (b) fundamentally change how RL methods function. These insights show the difficulty, and importance, of attributing performance gains in deep reinforcement learning.

A Closer Look at Deep Policy Gradients

- Andrew Ilyas, Logan Engstrom, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, Aleksander Madry
- abstract@open-review(Talk): We study how the behavior of deep policy gradient algorithms reflects the conceptual framework motivating their development. To this end, we propose a fine-grained analysis of state-of-the-art methods based on key elements of this framework: gradient estimation, value prediction, and optimization landscapes. Our results show that the behavior of deep policy gradient algorithms often deviates from what their motivating framework would predict: surrogate rewards do not match the true reward landscape, learned value estimators fail to fit the true value function, and gradient estimates poorly correlate with the "true" gradient. The mismatch between predicted and empirical behavior we uncover highlights our poor understanding of current methods, and indicates the need to move beyond current benchmark-centric evaluation methods.

Plug and Play Language Models: A Simple Approach to Controlled Text Generation

- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, Rosanne Liu
- abstract@open-review(Poster): Large transformer-based language models (LMs) trained on huge text corpora have shown unparalleled generation capabilities. However, controlling attributes of the generated language (e.g. switching topic or sentiment) is difficult without modifying the model architecture or fine-tuning on attribute-specific data and entailing the significant cost of retraining. We propose a simple alternative: the Plug and Play Language Model (PPLM) for controllable language generation, which combines a pretrained LM with one or more simple attribute classifiers that guide text generation without any further training of the LM. In the canonical scenario we present, the attribute models are simple classifiers consisting of a user-specified bag of words or a single learned layer with 100,000 times fewer parameters than the LM. Sampling entails a forward and backward pass in which gradients from the attribute model push the LM's hidden activations and thus guide the generation. Model samples demonstrate control over a range

of topics and sentiment styles, and extensive automated and human annotated evaluations show attribute alignment and fluency. PPLMs are flexible in that any combination of differentiable attribute models may be used to steer text generation, which will allow for diverse and creative applications beyond the examples given in this paper.

Understanding and Robustifying Differentiable Architecture Search

- Arber Zela, Thomas Elsken, Tonmoy Saikia, Yassine Marrakchi, Thomas Brox, Frank Hutter
- abstract@open-review(Talk): Differentiable Architecture Search (DARTS) has attracted a lot of attention due to its simplicity and small search costs achieved by a continuous relaxation and an approximation of the resulting bi-level optimization problem. However, DARTS does not work robustly for new problems: we identify a wide range of search spaces for which DARTS yields degenerate architectures with very poor test performance. We study this failure mode and show that, while DARTS successfully minimizes validation loss, the found solutions generalize poorly when they coincide with high validation loss curvature in the architecture space. We show that by adding one of various types of regularization we can robustify DARTS to find solutions with less curvature and better generalization properties. Based on these observations, we propose several simple variations of DARTS that perform substantially more robustly in practice. Our observations are robust across five search spaces on three image classification tasks and also hold for the very different domains of disparity estimation (a dense regression task) and language modelling.

Rethinking the Hyperparameters for Fine-tuning

- Hao Li, Pratik Chaudhari, Hao Yang, Michael Lam, Avinash Ravichandran, Rahul Bhotika, Stefano Soatto
- abstract@open-review(Poster): Fine-tuning from pre-trained ImageNet models has become the de-facto standard for various computer vision tasks. Current practices for fine-tuning typically involve selecting an ad-hoc choice of hyperparameters and keeping them fixed to values normally used for training from scratch. This paper re-examines several common practices of setting hyperparameters for fine-tuning. Our findings are based on extensive empirical evaluation for fine-tuning on various transfer learning benchmarks. (1) While prior works have thoroughly investigated learning rate and batch size, momentum for fine-tuning is a relatively unexplored parameter. We find that the value of momentum also affects fine-tuning performance and connect it with previous theoretical findings. (2) Optimal hyperparameters for fine-tuning, in particular, the effective learning rate, are not only dataset dependent but also sensitive to the similarity between the source domain and target domain. This is in contrast to hyperparameters for training from scratch. (3) Reference-based regularization that keeps models close to the initial model does not necessarily apply for "dissimilar" datasets. Our findings challenge common practices of fine-tuning and encourages deep learning practitioners to rethink the hyperparameters for fine-tuning.

Query2box: Reasoning over Knowledge Graphs in Vector Space Using Box Embeddings

- Hongyu Ren, *Weihua Hu*, Jure Leskovec
- abstract@open-review(Poster): Answering complex logical queries on large-scale incomplete knowledge graphs (KGs) is a fundamental yet challenging task. Recently, a promising approach to this problem has been to embed KG entities as well as the query into a vector space such that entities that answer the query are embedded close to the query. However, prior work models queries as single points in the vector space, which is problematic because a complex query represents a potentially large set of its answer entities, but it is unclear how such a set can be represented as a single point. Furthermore, prior work can only handle queries that use conjunctions ($\wedge$) and existential quantifiers ($\exists$). Handling queries with logical disjunctions ($\vee$) remains an open problem. Here we propose query2box, an embedding-based framework for reasoning over arbitrary queries with $\wedge$, $\vee$, and $\exists$ operators in massive and incomplete KGs. Our main insight is that queries can be embedded as boxes (i.e., hyper-rectangles), where a set of points inside the box corresponds to a set of answer entities of the query. We show that conjunctions can be naturally represented as intersections of boxes and also prove a negative result that handling disjunctions would require embedding with dimension proportional to the number of KG entities. However, we show that by transforming queries into a Disjunctive Normal Form, query2box is capable of handling arbitrary logical queries with $\wedge$, $\vee$, $\exists$ in a scalable manner. We demonstrate the effectiveness of query2box on two large KGs and show that query2box achieves up to 25% relative improvement over the state of the art.

Implementing Inductive bias for different navigation tasks through diverse RNN attrractors

- Tie XU, Omri Barak
- abstract@open-review(Poster): Navigation is crucial for animal behavior and is assumed to require an internal representation of the external environment, termed a cognitive map. The precise form of this representation is often considered to be a metric representation of space. An internal representation, however, is judged by its contribution to performance on a given task, and may thus vary between different types of navigation tasks. Here we train a recurrent neural network that controls an agent performing several navigation tasks in a simple environment. To focus on internal representations, we split learning into a task-agnostic pre-training stage that modifies internal connectivity and a task-specific Q learning stage that controls the network's output. We show that pre-training shapes the attractor landscape of the networks, leading to either a continuous attractor, discrete attractors or a disordered state. These structures induce bias onto the Q-Learning phase, leading to a performance pattern across the tasks corresponding to metric and topological regularities. Our results show that, in recurrent networks, inductive bias takes the form of attractor landscapes -- which can be shaped by pre-training and analyzed using dynamical systems methods. Furthermore, we demonstrate that non-metric representations are useful for navigation tasks.

PCMC-Net: Feature-based Pairwise Choice Markov Chains

- Alix Lhéritier
- abstract@open-review(Poster): Pairwise Choice Markov Chains (PCMC) have been recently introduced to overcome limitations of choice models based on traditional axioms unable to express empirical observations from modern behavior economics like context effects occurring when a choice between two options is altered by adding a third alternative. The inference approach that estimates the transition rates between each possible pair of alternatives via maximum likelihood suffers when the examples of each alternative are scarce and is inappropriate when new alternatives can be observed at test time. In this work, we propose an amortized inference approach for PCMC by embedding its definition into a neural network that represents transition rates as a function of the alternatives' and individual's features. We apply our construction to the complex case of airline itinerary booking where singletons are common (due to varying prices and individual-specific itineraries), and context effects and behaviors strongly dependent on market segments are observed. Experiments show our network significantly outperforming, in terms of prediction accuracy and logarithmic loss, feature engineered standard and latent class Multinomial Logit models as well as recent machine learning approaches.

Multi-Agent Interactions Modeling with Correlated Policies

- Minghuan Liu, Ming Zhou, Weinan Zhang, Yuzheng Zhuang, Jun Wang, Wulong Liu, Yong Yu
- abstract@open-review(Poster): In multi-agent systems, complex interacting behaviors arise due to the high correlations among agents. However, previous work on modeling multi-agent interactions from demonstrations is primarily constrained by assuming the independence among policies and their reward structures. In this paper, we cast the multi-agent interactions modeling problem into a multi-agent imitation learning framework with explicit modeling of correlated policies by approximating opponents' policies, which can recover agents' policies that can regenerate similar interactions. Consequently, we develop a Decentralized Adversarial Imitation Learning algorithm with Correlated policies (CoDAIL), which allows for decentralized training and execution. Various experiments demonstrate that CoDAIL can better regenerate complex interactions close to the demonstrators and outperforms state-of-the-art multi-agent imitation learning methods. Our code is available at \url{https://github.com/apexrl/CoDAIL}.

Once-for-All: Train One Network and Specialize it for Efficient Deployment

- Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, Song Han
- abstract@open-review(Poster): We address the challenging problem of efficient inference across many devices and resource constraints, especially on edge devices. Conventional approaches either manually design or use neural architecture search (NAS) to find a specialized neural network and train it from scratch for each case, which is computationally prohibitive (causing CO_2 emission as much as 5 cars' lifetime) thus unscalable. In this work, we propose to train a once-for-all (OFA) network that supports diverse architectural settings by decoupling training and search, to reduce the cost. We can quickly get a specialized sub-network by selecting from the OFA network without additional training. To efficiently train OFA networks, we also propose a novel progressive shrinking algorithm, a generalized pruning method that reduces the model size across many more dimensions than pruning (depth, width, kernel size, and resolution). It can obtain a surprisingly large number of sub-networks ($> 10^{19}$) that can fit different hardware platforms and latency constraints while maintaining the same level of accuracy as training independently. On diverse edge devices, OFA consistently outperforms state-of-the-art (SOTA) NAS methods (up to 4.0% ImageNet top1 accuracy improvement over MobileNetV3, or same accuracy but 1.5x faster than MobileNetV3, 2.6x faster than EfficientNet w.r.t measured latency) while reducing many orders of magnitude GPU hours and CO_2 emission. In particular, OFA achieves a new SOTA 80.0% ImageNet top-1 accuracy under the mobile setting ($<$600M MACs). OFA is the winning solution for the 3rd Low Power Computer Vision Challenge (LPCVC), DSP classification track and the 4th LPCVC, both classification track and detection track. Code and 50 pre-trained models (for many devices & many latency constraints) are released at https://github.com/mit-han-lab/once-for-all.

Generalized Convolutional Forest Networks for Domain Generalization and Visual Recognition

- Jongbin Ryu, Gitaek Kwon, Ming-Hsuan Yang, Jongwoo Lim
- abstract@open-review(Poster): When constructing random forests, it is of prime importance to ensure high accuracy and low correlation of individual tree classifiers for good performance. Nevertheless, it is typically difficult for existing random forest methods to strike a good balance between these conflicting factors. In this work, we propose a generalized convolutional forest networks to learn a feature space to maximize the strength of individual tree classifiers while minimizing the respective correlation. The feature space is iteratively constructed by a probabilistic triplet sampling method based on the distribution obtained from the splits of the random forest. The sampling process is designed to pull the data of the same label together for higher strength and push away the data frequently falling to the same leaf nodes. We perform extensive experiments on five image classification and two domain generalization datasets with ResNet-50 and DenseNet-161 backbone networks. Experimental results show that the proposed algorithm performs favorably against state-of-the-art methods.

SNODE: Spectral Discretization of Neural ODEs for System Identification

- Alessio Quaglino, Marco Gallieri, Jonathan Masci, Jan Koutník
- abstract@open-review(Poster): This paper proposes the use of spectral element methods \citep{canuto_spectral_1988} for fast and accurate training of Neural Ordinary Differential Equations (ODE-Nets; \citealp{Chen2018NeuralOD}) for system identification. This is achieved by expressing their dynamics as a truncated series of Legendre polynomials. The series coefficients, as well as the network weights, are computed by minimizing the weighted sum of the loss function and the violation of the ODE-Net dynamics. The problem is solved by coordinate descent that alternately minimizes, with respect to the coefficients and the weights, two unconstrained sub-problems using standard backpropagation and gradient methods. The resulting optimization scheme is fully time-parallel and results in a low memory footprint. Experimental comparison to standard methods, such as backpropagation through explicit solvers and the adjoint technique \citep{Chen2018NeuralOD}, on training surrogate models of small and medium-scale dynamical systems shows that it is at least one order of magnitude faster at reaching a comparable value of the loss function. The corresponding testing MSE is one order of magnitude smaller as well, suggesting generalization capabilities increase.

Guiding Program Synthesis by Learning to Generate Examples

- Larissa Laich, Pavol Bielik, Martin Vechev
- abstract@open-review(Poster): A key challenge of existing program synthesizers is ensuring that the synthesized program generalizes well. This can be difficult to achieve as the specification provided by the end user is often limited, containing as few as one or two input-output examples. In this paper we address this challenge via an iterative approach that finds ambiguities in the provided specification and learns to resolve these by generating additional input-output examples. The main insight is to reduce the problem of selecting which program generalizes well to the simpler task of deciding which output is correct. As a result, to train our probabilistic models, we can take advantage of the large amounts of data in the form of program outputs, which are often much easier to obtain than the corresponding ground-truth programs.

Fast Neural Network Adaptation via Parameter Remapping and Architecture Search

- Jiemin Fang, *Yuzhu Sun*, Kangjian Peng*, Qian Zhang, Yuan Li, Wenyu Liu, Xinggang Wang
- abstract@open-review(Poster): Deep neural networks achieve remarkable performance in many computer vision tasks. Most state-of-the-art~(SOTA) semantic segmentation and object detection approaches reuse neural network architectures designed for image classification as the backbone, commonly pre-trained on ImageNet. However, performance gains can be achieved by designing network architectures specifically for detection and segmentation, as shown by recent neural architecture search (NAS) research for detection and segmentation. One major challenge though, is that ImageNet pre-training of the search space representation (a.k.a. super network) or the searched networks incurs huge computational cost. In this paper, we propose a Fast Neural Network Adaptation (FNA) method, which can adapt both the architecture and parameters of a seed network (e.g. a high performing manually designed backbone) to become a network with different depth, width, or kernels via a Parameter Remapping technique, making it possible to utilize NAS for detection/segmentation tasks a lot more efficiently. In our experiments, we conduct FNA on MobileNetV2 to obtain new networks for both segmentation and detection that clearly out-perform existing networks designed both manually and by NAS. The total computation cost of FNA is significantly less than SOTA segmentation/detection NAS approaches: 1737$\times$ less than DPC, 6.8$\times$ less than Auto-DeepLab and 7.4$\times$ less than DetNAS. The code is available at https://github.com/JaminFong/FNA .

Exploratory Not Explanatory: Counterfactual Analysis of Saliency Maps for Deep Reinforcement Learning

- Akanksha Atrey, Kaleigh Clary, David Jensen
- abstract@open-review(Poster): Saliency maps are frequently used to support explanations of the behavior of deep reinforcement learning (RL) agents. However, a review of how saliency maps are used in practice indicates that the derived explanations are often unfalsifiable and can be highly subjective. We introduce an empirical approach grounded in counterfactual reasoning to test the hypotheses generated from saliency maps and assess the degree to which they correspond to the semantics of RL environments. We use Atari games, a common benchmark for deep RL, to evaluate three types of saliency maps. Our results show the extent to which existing claims about Atari games can be evaluated and suggest that saliency maps are best viewed as an exploratory tool rather than an explanatory tool.

Meta-Learning Deep Energy-Based Memory Models

- Sergey Bartunov, Jack Rae, Simon Osindero, Timothy Lillicrap

- abstract@open-review(Poster): We study the problem of learning an associative memory model -- a system which is able to retrieve a remembered pattern based on its distorted or incomplete version. Attractor networks provide a sound model of associative memory: patterns are stored as attractors of the network dynamics and associative retrieval is performed by running the dynamics starting from a query pattern until it converges to an attractor. In such models the dynamics are often implemented as an optimization procedure that minimizes an energy function, such as in the classical Hopfield network. In general it is difficult to derive a writing rule for a given dynamics and energy that is both compressive and fast. Thus, most research in energy-based memory has been limited either to tractable energy models not expressive enough to handle complex high-dimensional objects such as natural images, or to models that do not offer fast writing. We present a novel meta-learning approach to energy-based memory models (EBMM) that allows one to use an arbitrary neural architecture as an energy model and quickly store patterns in its weights. We demonstrate experimentally that our EBMM approach can build compressed memories for synthetic and natural data, and is capable of associative retrieval that outperforms existing memory systems in terms of the reconstruction error and compression rate.

Graph Convolutional Reinforcement Learning

- Jiechuan Jiang, Chen Dun, Tiejun Huang, Zongqing Lu
- abstract@open-review(Poster): Learning to cooperate is crucially important in multi-agent environments. The key is to understand the mutual interplay between agents. However, multi-agent environments are highly dynamic, where agents keep moving and their neighbors change quickly. This makes it hard to learn abstract representations of mutual interplay between agents. To tackle these difficulties, we propose graph convolutional reinforcement learning, where graph convolution adapts to the dynamics of the underlying graph of the multi-agent environment, and relation kernels capture the interplay between agents by their relation representations. Latent features produced by convolutional layers from gradually increased receptive fields are exploited to learn cooperation, and cooperation is further improved by temporal relation regularization for consistency. Empirically, we show that our method substantially outperforms existing methods in a variety of cooperative scenarios.

Kernelized Wasserstein Natural Gradient

- M Arbel, A Gretton, W Li, G Montufar
- abstract@open-review(Spotlight): Many machine learning problems can be expressed as the optimization of some cost functional over a parametric family of probability distributions. It is often beneficial to solve such optimization problems using natural gradient methods. These methods are invariant to the parametrization of the family, and thus can yield more effective optimization. Unfortunately, computing the natural gradient is challenging as it requires inverting a high dimensional matrix at each iteration. We propose a general framework to approximate the natural gradient for the Wasserstein metric, by leveraging a dual formulation of the metric restricted to a Reproducing Kernel Hilbert Space. Our approach leads to an estimator for gradient direction that can trade-off accuracy and computational cost, with theoretical guarantees. We verify its accuracy on simple examples, and show the advantage of using such an estimator in classification tasks on \texttt{Cifar10} and \texttt{Cifar100} empirically.

The Curious Case of Neural Text Degeneration

- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, Yejin Choi
- abstract@open-review(Poster): Despite considerable advances in neural language modeling, it remains an open question what the best decoding strategy is for text generation from a language model (e.g. to generate a story). The counter-intuitive empirical observation is that even though the use of likelihood as training objective leads to high quality models for a broad range of language understanding tasks, maximization-based decoding methods such as beam search lead to degeneration — output text that is bland, incoherent, or gets stuck in repetitive loops.

To address this we propose Nucleus Sampling, a simple but effective method to draw considerably higher quality text out of neural language models than previous decoding strategies. Our approach avoids text degeneration by truncating the unreliable tail of the probability distribution, sampling from the dynamic nucleus of tokens containing the vast majority of the probability mass.

To properly examine current maximization-based and stochastic decoding methods, we compare generations from each of these methods to the distribution of human text along several axes such as likelihood, diversity, and repetition. Our results show that (1) maximization is an inappropriate decoding objective for open-ended text generation, (2) the probability distributions of the best current language models have an unreliable tail which needs to be truncated during generation and (3) Nucleus Sampling is currently the best available decoding strategy for generating long-form text that is both high-quality — as measured by human evaluation — and as diverse as human-written text.

Multilingual Alignment of Contextual Word Representations

- Steven Cao, Nikita Kitaev, Dan Klein
- abstract@open-review(Poster): We propose procedures for evaluating and strengthening contextual embedding alignment and show that they are useful in analyzing and improving multilingual BERT. In particular, after our proposed alignment procedure, BERT exhibits significantly improved zero-shot performance on XNLI compared to the base model, remarkably matching pseudo-fully-supervised translate-train models for Bulgarian and Greek. Further, to measure the degree of alignment, we introduce a contextual version of word retrieval and show that it correlates well with downstream zero-shot transfer. Using this word retrieval task, we also analyze BERT and find that it exhibits systematic deficiencies, e.g. worse alignment for open-class parts-of-speech and word pairs written in different scripts, that are corrected by the alignment procedure. These results support contextual alignment as a useful concept for understanding large multilingual pre-trained models.

The Gambler's Problem and Beyond

- Baoxiang Wang, Shuai Li, Jiajin Li, Siu On Chan
- abstract@open-review(Poster): We analyze the Gambler's problem, a simple reinforcement learning problem where the gambler has the chance to double or lose their bets until the target is reached. This is an early example introduced in the reinforcement learning textbook by Sutton and Barto (2018), where they mention an interesting pattern of the optimal value function with high-frequency components and repeating non-smooth points. It is however without further investigation. We provide the exact formula for the optimal value function for both the discrete and the continuous cases. Though simple as it might seem, the value function is pathological: fractal, self-similar, derivative taking either zero or infinity, not smooth on any interval, and not written as elementary functions. It is in fact one of the generalized Cantor functions, where it holds a complexity that has been uncharted thus far. Our analyses could lead insights into improving value function approximation, gradient-based algorithms, and Q-learning, in real applications and implementations.

GraphAF: a Flow-based Autoregressive Model for Molecular Graph Generation

- Chence Shi, *Minkai Xu*, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, Jian Tang
- abstract@open-review(Poster): Molecular graph generation is a fundamental problem for drug discovery and has been attracting growing attention. The problem is challenging since it requires not only generating chemically valid molecular structures but also optimizing their chemical properties in the meantime. Inspired by the recent progress in deep generative models, in this paper we propose a flow-based autoregressive model for graph generation called GraphAF. GraphAF combines the advantages of both autoregressive and flow-based approaches and enjoys: (1) high model flexibility for data density estimation; (2) efficient parallel computation for training; (3) an iterative sampling process, which allows leveraging chemical domain knowledge for valency checking. Experimental results show that GraphAF is able to generate 68\% chemically valid molecules even without chemical knowledge

rules and 100\% valid molecules with chemical rules. The training process of GraphAF is two times faster than the existing state-of-the-art approach GCPN. After fine-tuning the model for goal-directed property optimization with reinforcement learning, GraphAF achieves state-of-the-art performance on both chemical property optimization and constrained property optimization.

Double Neural Counterfactual Regret Minimization

- Hui Li, Kailiang Hu, Shaohua Zhang, Yuan Qi, Le Song
- abstract@open-review(Poster): Counterfactual regret minimization (CFR) is a fundamental and effective technique for solving Imperfect Information Games (IIG). However, the original CFR algorithm only works for discrete states and action spaces, and the resulting strategy is maintained as a tabular representation. Such tabular representation limits the method from being directly applied to large games. In this paper, we propose a double neural representation for the IIGs, where one neural network represents the cumulative regret, and the other represents the average strategy. Such neural representations allow us to avoid manual game abstraction and carry out end-to-end optimization. To make the learning efficient, we also developed several novel techniques including a robust sampling method and a mini-batch Monte Carlo Counterfactual Regret Minimization (MCCFR) method, which may be of independent interests. Empirically, on games tractable to tabular approaches, neural strategies trained with our algorithm converge comparably to their tabular counterparts, and significantly outperform those based on deep reinforcement learning. On extremely large games with billions of decision nodes, our approach achieved strong performance while using hundreds of times less memory than the tabular CFR. On head-to-head matches of hands-up no-limit texas hold'em, our neural agent beat the strong agent ABS-CFR by 9.8 ± 4.1 chips per game. It's a successful application of neural CFR in large games.

Neural Policy Gradient Methods: Global Optimality and Rates of Convergence

- Lingxiao Wang, Qi Cai, Zhuoran Yang, Zhaoran Wang
- abstract@open-review(Poster): Policy gradient methods with actor-critic schemes demonstrate tremendous empirical successes, especially when the actors and critics are parameterized by neural networks. However, it remains less clear whether such "neural" policy gradient methods converge to globally optimal policies and whether they even converge at all. We answer both the questions affirmatively in the overparameterized regime. In detail, we prove that neural natural policy gradient converges to a globally optimal policy at a sublinear rate. Also, we show that neural vanilla policy gradient converges sublinearly to a stationary point. Meanwhile, by relating the suboptimality of the stationary points to the~representation power of neural actor and critic classes, we prove the global optimality of all stationary points under mild regularity conditions. Particularly, we show that a key to the global optimality and convergence is the "compatibility" between the actor and critic, which is ensured by sharing neural architectures and random initializations across the actor and critic. To the best of our knowledge, our analysis establishes the first global optimality and convergence guarantees for neural policy gradient methods.

Triple Wins: Boosting Accuracy, Robustness and Efficiency Together by Enabling Input-Adaptive Inference

- Ting-Kuei Hu, Tianlong Chen, Haotao Wang, Zhangyang Wang
- abstract@open-review(Poster): Deep networks were recently suggested to face the odds between accuracy (on clean natural images) and robustness (on adversarially perturbed images) (Tsipras et al., 2019). Such a dilemma is shown to be rooted in the inherently higher sample complexity (Schmidt et al., 2018) and/or model capacity (Nakkiran, 2019), for learning a high-accuracy and robust classifier. In view of that, give a classification task, growing the model capacity appears to help draw a win-win between accuracy and robustness, yet at the expense of model size and latency, therefore posing challenges for resource-constrained applications. Is it possible to co-design model accuracy, robustness and efficiency to achieve their triple wins? This paper studies multi-exit networks associated with input-adaptive efficient inference, showing their strong promise in achieving a “sweet point" in co-optimizing model accuracy, robustness, and efficiency. Our proposed solution, dubbed Robust Dynamic Inference Networks (RDI-Nets), allows for each input (either clean or adversarial) to adaptively choose one of the multiple output layers (early branches or the final one) to output its prediction. That multi-loss adaptivity adds new variations and flexibility to adversarial attacks and defenses, on which we present a systematical investigation. We show experimentally that by equipping existing backbones with such robust adaptive inference, the resulting RDI-Nets can achieve better accuracy and robustness, yet with over 30% computational savings, compared to the defended original models.

Dynamic Sparse Training: Find Efficient Sparse Network From Scratch With Trainable Masked Layers

- Junjie LIU, Zhe XU, Runbin SHI, Ray C. C. Cheung, Hayden K.H. So
- abstract@open-review(Poster): We present a novel network pruning algorithm called Dynamic Sparse Training that can jointly find the optimal network parameters and sparse network structure in a unified optimization process with trainable pruning thresholds. These thresholds can have fine-grained layer-wise adjustments dynamically via backpropagation. We demonstrate that our dynamic sparse training algorithm can easily train very sparse neural network models with little performance loss using the same training epochs as dense models. Dynamic Sparse Training achieves prior art performance compared with other sparse training algorithms on various network architectures. Additionally, we have several surprising observations that provide strong evidence to the effectiveness and efficiency of our algorithm. These observations reveal the underlying problems of traditional three-stage pruning algorithms and present the potential guidance provided by our algorithm to the design of more compact network architectures.

Beyond Linearization: On Quadratic and Higher-Order Approximation of Wide Neural Networks

- Yu Bai, Jason D. Lee
- abstract@open-review(Poster): Recent theoretical work has established connections between over-parametrized neural networks and linearized models governed by the Neural Tangent Kernels (NTKs). NTK theory leads to concrete convergence and generalization results, yet the empirical performance of neural networks are observed to exceed their linearized models, suggesting insufficiency of this theory. Towards closing this gap, we investigate the training of over-parametrized neural networks that are beyond the NTK regime yet still governed by the Taylor expansion of the network. We bring forward the idea of randomizing the neural networks, which allows them to escape their NTK and couple with quadratic models. We show that the optimization landscape of randomized two-layer networks are nice and amenable to escaping-saddle algorithms. We prove concrete generalization and expressivity results on these randomized networks, which lead to sample complexity bounds (of learning certain simple functions) that match the NTK and can in addition be better by a dimension factor when mild distributional assumptions are present. We demonstrate that our randomization technique can be generalized systematically beyond the quadratic case, by using it to find networks that are coupled with higher-order terms in their Taylor series.

Deep Graph Matching Consensus

- Matthias Fey, Jan E. Lenssen, Christopher Morris, Jonathan Masci, Nils M. Kriege
- abstract@open-review(Poster): This work presents a two-stage neural architecture for learning and refining structural correspondences between graphs. First, we use localized node embeddings computed by a graph neural network to obtain an initial ranking of soft correspondences between nodes. Secondly, we employ synchronous message passing networks to iteratively re-rank the soft correspondences to reach a matching consensus in local neighborhoods between graphs. We show, theoretically and empirically, that our message passing scheme computes a well-founded measure of consensus for corresponding neighborhoods, which is then used to guide the iterative re-ranking process. Our purely local and sparsity-aware architecture scales well to large, real-world inputs while still being able to recover global correspondences consistently. We demonstrate the practical effectiveness of our method on real-world tasks from the fields of computer vision and entity alignment between knowledge graphs, on which we improve upon the current state-of-the-art.

Self-Supervised Learning of Appliance Usage

- Chen-Yu Hsu, Abbas Zeitoun, Guang-He Lee, Dina Katabi, Tommi Jaakkola
- abstract@open-review(Poster): Learning home appliance usage is important for understanding people's activities and optimizing energy consumption. The problem is modeled as an event detection task, where the objective is to learn when a user turns an appliance on, and which appliance it is (microwave, hair dryer, etc.). Ideally, we would like to solve the problem in an unsupervised way so that the method can be applied to new homes and new appliances without any labels. To this end, we introduce a new deep learning model that takes input from two home sensors: 1) a smart electricity meter that outputs the total energy consumed by the home as a function of time, and 2) a motion sensor that outputs the locations of the residents over time. The model learns the distribution of the residents' locations conditioned on the home energy signal. We show that this cross-modal prediction task allows us to detect when a particular appliance is used, and the location of the appliance in the home, all in a self-supervised manner, without any labeled data.

Quantum Algorithms for Deep Convolutional Neural Networks

- Iordanis Kerenidis, Jonas Landman, Anupam Prakash
- abstract@open-review(Poster): Quantum computing is a powerful computational paradigm with applications in several fields, including machine learning. In the last decade, deep learning, and in particular Convolutional Neural Networks (CNN), have become essential for applications in signal processing and image recognition. Quantum deep learning, however, remains a challenging problem, as it is difficult to implement non linearities with quantum unitaries. In this paper we propose a quantum algorithm for evaluating and training deep convolutional neural networks with potential speedups over classical CNNs for both the forward and backward passes. The quantum CNN (QCNN) reproduces completely the outputs of the classical CNN and allows for non linearities and pooling operations. The QCNN is in particular interesting for deep networks and could allow new frontiers in the image recognition domain, by allowing for many more convolution kernels, larger kernels, high dimensional inputs and high depth input channels. We also present numerical simulations for the classification of the MNIST dataset to provide practical evidence for the efficiency of the QCNN.

Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds

- Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, Alekh Agarwal
- abstract@open-review(Talk): We design a new algorithm for batch active learning with deep neural network models. Our algorithm, Batch Active learning by Diverse Gradient Embeddings (BADGE), samples groups of points that are disparate and high-magnitude when represented in a hallucinated gradient space, a strategy designed to incorporate both predictive uncertainty and sample diversity into every selected batch. Crucially, BADGE trades off between diversity and uncertainty without requiring any hand-tuned hyperparameters. While other approaches sometimes succeed for particular batch sizes or architectures, BADGE consistently performs as well or better, making it a useful option for real world active learning problems.

Understanding l4-based Dictionary Learning: Interpretation, Stability, and Robustness

- Yuexiang Zhai, Hermish Mehta, Zhengyuan Zhou, Yi Ma
- abstract@open-review(Poster): Recently, the ℓ^4-norm maximization has been proposed to solve the sparse dictionary learning (SDL) problem. The simple MSP (matching, stretching, and projection) algorithm proposed by \cite{zhai2019a} has proved surprisingly efficient and effective. This paper aims to better understand this algorithm from its strong geometric and statistical connections with the classic PCA and ICA, as well as their associated fixed-point style algorithms. Such connections provide a unified way of viewing problems that pursue {\em principal}, {\em independent}, or {\em sparse} components of high-dimensional data. Our studies reveal additional good properties of ℓ^4-maximization: not only is the MSP algorithm for sparse coding insensitive to small noise, but it is also robust to outliers and resilient to sparse corruptions. We provide statistical justification for such inherently nice properties. To corroborate the theoretical analysis, we also provide extensive and compelling experimental evidence with both synthetic data and real images.

Symplectic ODE-Net: Learning Hamiltonian Dynamics with Control

- Yaofeng Desmond Zhong, Biswadip Dey, Amit Chakraborty
- abstract@open-review(Poster): In this paper, we introduce Symplectic ODE-Net (SymODEN), a deep learning framework which can infer the dynamics of a physical system, given by an ordinary differential equation (ODE), from observed state trajectories. To achieve better generalization with fewer training samples, SymODEN incorporates appropriate inductive bias by designing the associated computation graph in a physics-informed manner. In particular, we enforce Hamiltonian dynamics with control to learn the underlying dynamics in a transparent way, which can then be leveraged to draw insight about relevant physical aspects of the system, such as mass and potential energy. In addition, we propose a parametrization which can enforce this Hamiltonian formalism even when the generalized coordinate data is embedded in a high-dimensional space or we can only access velocity data instead of generalized momentum. This framework, by offering interpretable, physically-consistent models for physical systems, opens up new possibilities for synthesizing model-based control strategies.

FreeLB: Enhanced Adversarial Training for Natural Language Understanding

- Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, Jingjing Liu
- abstract@open-review(Spotlight): Adversarial training, which minimizes the maximal risk for label-preserving input perturbations, has proved to be effective for improving the generalization of language models. In this work, we propose a novel adversarial training algorithm, FreeLB, that promotes higher invariance in the embedding space, by adding adversarial perturbations to word embeddings and minimizing the resultant adversarial risk inside different regions around input samples. To validate the effectiveness of the proposed approach, we apply it to Transformer-based models for natural language understanding and commonsense reasoning tasks. Experiments on the GLUE benchmark show that when applied only to the finetuning stage, it is able to improve the overall test scores of BERT-base model from 78.3 to 79.4, and RoBERTa-large model from 88.5 to 88.8. In addition, the proposed approach achieves state-of-the-art single-model test accuracies of 85.44% and 67.75% on ARC-Easy and ARC-Challenge. Experiments on CommonsenseQA benchmark further demonstrate that FreeLB can be generalized and boost the performance of RoBERTa-large model on other tasks as well.

Behaviour Suite for Reinforcement Learning

- Ian Osband, Yotam Doron, Matteo Hessel, John Aslanides, Eren Sezener, Andre Saraiva, Katrina McKinney, Tor Lattimore, Csaba Szepesvari, Satinder Singh, Benjamin Van Roy, Richard Sutton, David Silver, Hado Van Hasselt
- abstract@open-review(Spotlight): This paper introduces the Behaviour Suite for Reinforcement Learning, or bsuite for short. bsuite is a collection of carefully-designed experiments that investigate core capabilities of reinforcement learning (RL) agents with two objectives. First, to collect clear, informative and scalable problems that capture key issues in the design of general and efficient learning algorithms. Second, to study agent behaviour through their performance on these shared benchmarks. To complement this effort, we open source this http URL, which automates evaluation and analysis of any agent on bsuite. This library facilitates reproducible and accessible research on the core issues in RL, and ultimately the design of superior learning algorithms. Our code is Python, and easy to use within existing projects. We include examples with OpenAI Baselines, Dopamine as well as new reference implementations. Going forward, we hope to incorporate more excellent experiments from the research community, and commit to a periodic review of bsuite from a committee of prominent researchers.

Strategies for Pre-training Graph Neural Networks

- Weihua Hu, *Bowen Liu*, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, Jure Leskovec
- abstract@open-review(Spotlight): Many applications of machine learning require a model to make accurate pre-dictions on test examples that are distributionally different from training ones, while task-specific labels are scarce during training. An effective approach to this challenge is to pre-train a model on related tasks where data is abundant, and then fine-tune it on a downstream task of interest. While pre-training has been effective in many language and vision domains, it remains an open question how to effectively use pre-training on graph datasets. In this paper, we develop a new strategy and self-supervised methods for pre-training Graph Neural Networks (GNNs). The key to the success of our strategy is to pre-train an expressive GNN at the level of individual nodes as well as entire graphs so that the GNN can learn useful local and global representations simultaneously. We systematically study pre-training on multiple graph classification datasets. We find that naïve strategies, which pre-train GNNs at the level of either entire graphs or individual nodes, give limited improvement and can even lead to negative transfer on many downstream tasks. In contrast, our strategy avoids negative transfer and improves generalization significantly across downstream tasks, leading up to 9.4% absolute improvements in ROC-AUC over non-pre-trained models and achieving state-of-the-art performance for molecular property prediction and protein function prediction.

NAS-Bench-201: Extending the Scope of Reproducible Neural Architecture Search

- Xuanyi Dong, Yi Yang
- abstract@open-review(Spotlight): Neural architecture search (NAS) has achieved breakthrough success in a great number of applications in the past few years. It could be time to take a step back and analyze the good and bad aspects in the field of NAS. A variety of algorithms search architectures under different search space. These searched architectures are trained using different setups, e.g., hyper-parameters, data augmentation, regularization. This raises a comparability problem when comparing the performance of various NAS algorithms. NAS-Bench-101 has shown success to alleviate this problem. In this work, we propose an extension to NAS-Bench-101: NAS-Bench-201 with a different search space, results on multiple datasets, and more diagnostic information. NAS-Bench-201 has a fixed search space and provides a unified benchmark for almost any up-to-date NAS algorithms. The design of our search space is inspired by the one used in the most popular cell-based searching algorithms, where a cell is represented as a directed acyclic graph. Each edge here is associated with an operation selected from a predefined operation set. For it to be applicable for all NAS algorithms, the search space defined in NAS-Bench-201 includes all possible architectures generated by 4 nodes and 5 associated operation options, which results in 15,625 neural cell candidates in total. The training log using the same setup and the performance for each architecture candidate are provided for three datasets. This allows researchers to avoid unnecessary repetitive training for selected architecture and focus solely on the search algorithm itself. The training time saved for every architecture also largely improves the efficiency of most NAS algorithms and presents a more computational cost friendly NAS community for a broader range of researchers. We provide additional diagnostic information such as fine-grained loss and accuracy, which can give inspirations to new designs of NAS algorithms. In further support of the proposed NAS-Bench-102, we have analyzed it from many aspects and benchmarked 10 recent NAS algorithms, which verify its applicability.

Controlling generative models with continuous factors of variations

- Antoine Plumerault, Hervé Le Borgne, Céline Hudelot
- abstract@open-review(Poster): Recent deep generative models can provide photo-realistic images as well as visual or textual content embeddings useful to address various tasks of computer vision and natural language processing. Their usefulness is nevertheless often limited by the lack of control over the generative process or the poor understanding of the learned representation. To overcome these major issues, very recent works have shown the interest of studying the semantics of the latent space of generative models. In this paper, we propose to advance on the interpretability of the latent space of generative models by introducing a new method to find meaningful directions in the latent space of any generative model along which we can move to control precisely specific properties of the generated image like position or scale of the object in the image. Our method is weakly supervised and particularly well suited for the search of directions encoding simple transformations of the generated image, such as translation, zoom or color variations. We demonstrate the effectiveness of our method qualitatively and quantitatively, both for GANs and variational auto-encoders.

Emergent Tool Use From Multi-Agent Autocurricula

- Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, Igor Mordatch
- abstract@open-review(Spotlight): Through multi-agent competition, the simple objective of hide-and-seek, and standard reinforcement learning algorithms at scale, we find that agents create a self-supervised autocurriculum inducing multiple distinct rounds of emergent strategy, many of which require sophisticated tool use and coordination. We find clear evidence of six emergent phases in agent strategy in our environment, each of which creates a new pressure for the opposing team to adapt; for instance, agents learn to build multi-object shelters using moveable boxes which in turn leads to agents discovering that they can overcome obstacles using ramps. We further provide evidence that multi-agent competition may scale better with increasing environment complexity and leads to behavior that centers around far more human-relevant skills than other self-supervised reinforcement learning methods such as intrinsic motivation. Finally, we propose transfer and fine-tuning as a way to quantitatively evaluate targeted capabilities, and we compare hide-and-seek agents to both intrinsic motivation and random initialization baselines in a suite of domain-specific intelligence tests.

Simple and Effective Regularization Methods for Training on Noisily Labeled Data with Generalization Guarantee

- Wei Hu, Zhiyuan Li, Dingli Yu
- abstract@open-review(Poster): Over-parameterized deep neural networks trained by simple first-order methods are known to be able to fit any labeling of data. Such over-fitting ability hinders generalization when mislabeled training examples are present. On the other hand, simple regularization methods like early-stopping can often achieve highly nontrivial performance on clean test data in these scenarios, a phenomenon not theoretically understood. This paper proposes and analyzes two simple and intuitive regularization methods: (i) regularization by the distance between the network parameters to initialization, and (ii) adding a trainable auxiliary variable to the network output for each training example. Theoretically, we prove that gradient descent training with either of these two methods leads to a generalization guarantee on the clean data distribution despite being trained using noisy labels. Our generalization analysis relies on the connection between wide neural network and neural tangent kernel (NTK). The generalization bound is independent of the network size, and is comparable to the bound one can get when there is no label noise. Experimental results verify the effectiveness of these methods on noisily labeled datasets.

Unsupervised Clustering using Pseudo-semi-supervised Learning

- Divam Gupta, Ramachandran Ramjee, Nipun Kwatra, Muthian Sivathanu
- abstract@open-review(Poster): In this paper, we propose a framework that leverages semi-supervised models to improve unsupervised clustering performance. To leverage semi-supervised models, we first need to automatically generate labels, called pseudo-labels. We find that prior approaches for generating pseudo-labels hurt clustering performance because of their low accuracy. Instead, we use an ensemble of deep networks to construct a similarity graph, from which we extract high accuracy pseudo-labels. The approach of finding high quality pseudo-labels using ensembles and training the semi-supervised model is iterated, yielding continued improvement. We show that our approach outperforms state of the art clustering results for multiple image and text datasets. For example, we achieve 54.6% accuracy for CIFAR-10 and 43.9% for 20news, outperforming state of the art by 8-12% in absolute terms.

Geometric Analysis of Nonconvex Optimization Landscapes for Overcomplete Learning

- Qing Qu, Yuexiang Zhai, Xiao Li, Yuqian Zhang, Zhihui Zhu
- abstract@open-review(Talk): Learning overcomplete representations finds many applications in machine learning and data analytics. In the past decade, despite the empirical success of heuristic methods, theoretical understandings and explanations of these algorithms are still far from satisfactory. In this work, we provide new theoretical insights for several important representation learning problems: learning (i) sparsely used overcomplete dictionaries and (ii) convolutional dictionaries. We formulate these problems as ℓ^4-norm optimization problems over the sphere and study the geometric properties of their nonconvex optimization landscapes. For both problems, we show the nonconvex objective has benign (global) geometric structures, which enable the development of efficient optimization methods finding the target solutions. Finally, our theoretical results are justified by numerical simulations.

PairNorm: Tackling Oversmoothing in GNNs

- Lingxiao Zhao, Leman Akoglu
- abstract@open-review(Poster): The performance of graph neural nets (GNNs) is known to gradually decrease with increasing number of layers. This decay is partly attributed to oversmoothing, where repeated graph convolutions eventually make node embeddings indistinguishable. We take a closer look at two different interpretations, aiming to quantify oversmoothing. Our main contribution is PairNorm, a novel normalization layer that is based on a careful analysis of the graph convolution operator, which prevents all node embeddings from becoming too similar. What is more, PairNorm is fast, easy to implement without any change to network architecture nor any additional parameters, and is broadly applicable to any GNN. Experiments on real-world graphs demonstrate that PairNorm makes deeper GCN, GAT, and SGC models more robust against oversmoothing, and significantly boosts performance for a new problem setting that benefits from deeper GNNs. Code is available at https://github.com/LingxiaoShawn/PairNorm.

Black-box Off-policy Estimation for Infinite-Horizon Reinforcement Learning

- Ali Mousavi, Lihong Li, Qiang Liu, Denny Zhou
- abstract@open-review(Poster): Off-policy estimation for long-horizon problems is important in many real-life applications such as healthcare and robotics, where high-fidelity simulators may not be available and on-policy evaluation is expensive or impossible. Recently, \citet{liu18breaking} proposed an approach that avoids the curse of horizon suffered by typical importance-sampling-based methods. While showing promising results, this approach is limited in practice as it requires data being collected by a known behavior policy. In this work, we propose a novel approach that eliminates such limitations. In particular, we formulate the problem as solving for the fixed point of a "backward flow" operator and show that the fixed point solution gives the desired importance ratios of stationary distributions between the target and behavior policies. We analyze its asymptotic consistency and finite-sample generalization. Experiments on benchmarks verify the effectiveness of our proposed approach.

Empirical Studies on the Properties of Linear Regions in Deep Neural Networks

- Xiao Zhang, Dongrui Wu
- abstract@open-review(Poster): A deep neural networks (DNN) with piecewise linear activations can partition the input space into numerous small linear regions, where different linear functions are fitted. It is believed that the number of these regions represents the expressivity of a DNN. This paper provides a novel and meticulous perspective to look into DNNs: Instead of just counting the number of the linear regions, we study their local properties, such as the inspheres, the directions of the corresponding hyperplanes, the decision boundaries, and the relevance of the surrounding regions. We empirically observed that different optimization techniques lead to completely different linear regions, even though they result in similar classification accuracies. We hope our study can inspire the design of novel optimization techniques, and help discover and analyze the behaviors of DNNs.

SNOW: Subscribing to Knowledge via Channel Pooling for Transfer & Lifelong Learning of Convolutional Neural Networks

- Chungkuk Yoo, Bumsoo Kang, Minsik Cho
- abstract@open-review(Poster): SNOW is an efficient learning method to improve training/serving throughput as well as accuracy for transfer and lifelong learning of convolutional neural networks based on knowledge subscription. SNOW selects the top-K useful intermediate feature maps for a target task from a pre-trained and frozen source model through a novel channel pooling scheme, and utilizes them in the task-specific delta model. The source model is responsible for generating a large number of generic feature maps. Meanwhile, the delta model selectively subscribes to those feature maps and fuses them with its local ones to deliver high accuracy for the target task. Since a source model takes part in both training and serving of all target tasks in an inference-only mode, one source model can serve multiple delta models, enabling significant computation sharing. The sizes of such delta models are fractional of the source model, thus SNOW also provides model-size efficiency. Our experimental results show that SNOW offers a superior balance between accuracy and training/inference speed for various image classification tasks to the existing transfer and lifelong learning practices.

Smoothness and Stability in GANs

- Casey Chu, Kentaro Minami, Kenji Fukumizu
- abstract@open-review(Poster): Generative adversarial networks, or GANs, commonly display unstable behavior during training. In this work, we develop a principled theoretical framework for understanding the stability of various types of GANs. In particular, we derive conditions that guarantee eventual stationarity of the generator when it is trained with gradient descent, conditions that must be satisfied by the divergence that is minimized by the GAN and the generator's architecture. We find that existing GAN variants satisfy some, but not all, of these conditions. Using tools from convex analysis, optimal transport, and reproducing kernels, we construct a GAN that fulfills these conditions simultaneously. In the process, we explain and clarify the need for various existing GAN stabilization techniques, including Lipschitz constraints, gradient penalties, and smooth activation functions.

Adaptive Correlated Monte Carlo for Contextual Categorical Sequence Generation

- Xinjie Fan, Yizhe Zhang, Zhendong Wang, Mingyuan Zhou
- abstract@open-review(Poster): Sequence generation models are commonly refined with reinforcement learning over user-defined metrics. However, high gradient variance hinders the practical use of this method. To stabilize this method, we adapt to contextual generation of categorical sequences a policy gradient estimator, which evaluates a set of correlated Monte Carlo (MC) rollouts for variance control. Due to the correlation, the number of unique rollouts is random and adaptive to model uncertainty; those rollouts naturally become baselines for each other, and hence are combined to effectively reduce gradient variance. We also demonstrate the use of correlated MC rollouts for binary-tree softmax models, which reduce the high generation cost in large vocabulary scenarios by decomposing each categorical action into a sequence of binary actions. We evaluate our methods on both neural program synthesis and image captioning. The proposed methods yield lower gradient variance and consistent improvement over related baselines.

On Bonus Based Exploration Methods In The Arcade Learning Environment

- Adrien Ali Taiga, William Fedus, Marlos C. Machado, Aaron Courville, Marc G. Bellemare
- abstract@open-review(Poster): Research on exploration in reinforcement learning, as applied to Atari 2600 game-playing, has emphasized tackling difficult exploration problems such as Montezuma's Revenge (Bellemare et al., 2016). Recently, bonus-based exploration methods, which explore by augmenting the environment reward, have reached above-human average performance on such domains. In this paper we reassess popular bonus-based exploration methods within a common evaluation framework. We combine Rainbow (Hessel et al., 2018) with different exploration bonuses and evaluate its performance on Montezuma's Revenge, Bellemare et al.'s set of hard of exploration games with sparse rewards, and the whole Atari 2600 suite. We find

that while exploration bonuses lead to higher score on Montezuma's Revenge they do not provide meaningful gains over the simpler epsilon-greedy scheme. In fact, we find that methods that perform best on that game often underperform epsilon-greedy on easy exploration Atari 2600 games. We find that our conclusions remain valid even when hyperparameters are tuned for these easy-exploration games. Finally, we find that none of the methods surveyed benefit from additional training samples (1 billion frames, versus Rainbow's 200 million) on Bellemare et al.'s hard exploration games. Our results suggest that recent gains in Montezuma's Revenge may be better attributed to architecture change, rather than better exploration schemes; and that the real pace of progress in exploration research for Atari 2600 games may have been obfuscated by good results on a single domain.

A Theory of Usable Information under Computational Constraints

- Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, Stefano Ermon
- abstract@open-review(Talk): We propose a new framework for reasoning about information in complex systems. Our foundation is based on a variational extension of Shannon’s information theory that takes into account the modeling power and computational constraints of the observer. The resulting predictive V-information encompasses mutual information and other notions of informativeness such as the coefficient of determination. Unlike Shannon’s mutual information and in violation of the data processing inequality, V-information can be created through computation. This is consistent with deep neural networks extracting hierarchies of progressively more informative features in representation learning. Additionally, we show that by incorporating computational constraints, V-information can be reliably estimated from data even in high dimensions with PAC-style guarantees. Empirically, we demonstrate predictive V-information is more effective than mutual information for structure learning and fair representation learning. Codes are available at https://github.com/Newbeeer/V-information .

IMPACT: Importance Weighted Asynchronous Architectures with Clipped Target Networks

- Michael Luo, Jiahao Yao, Richard Liaw, Eric Liang, Ion Stoica
- abstract@open-review(Poster): The practical usage of reinforcement learning agents is often bottlenecked by the duration of training time. To accelerate training, practitioners often turn to distributed reinforcement learning architectures to parallelize and accelerate the training process. However, modern methods for scalable reinforcement learning (RL) often tradeoff between the throughput of samples that an RL agent can learn from (sample throughput) and the quality of learning from each sample (sample efficiency). In these scalable RL architectures, as one increases sample throughput (i.e. increasing parallelization in IMPALA (Espeholt et al., 2018)), sample efficiency drops significantly. To address this, we propose a new distributed reinforcement learning algorithm, IMPACT. IMPACT extends PPO with three changes: a target network for stabilizing the surrogate objective, a circular buffer, and truncated importance sampling. In discrete action-space environments, we show that IMPACT attains higher reward and, simultaneously, achieves up to 30% decrease in training wall-time than that of IMPALA. For continuous control environments, IMPACT trains faster than existing scalable agents while preserving the sample efficiency of synchronous PPO.

HiLLoC: lossless image compression with hierarchical latent variable models

- James Townsend, Thomas Bird, Julius Kunze, David Barber
- abstract@open-review(Poster): We make the following striking observation: fully convolutional VAE models trained on 32x32 ImageNet can generalize well, not just to 64x64 but also to far larger photographs, with no changes to the model. We use this property, applying fully convolutional models to lossless compression, demonstrating a method to scale the VAE-based 'Bits-Back with ANS' algorithm for lossless compression to large color photographs, and achieving state of the art for compression of full size ImageNet images. We release Craystack, an open source library for convenient prototyping of lossless compression using probabilistic models, along with full implementations of all of our compression results.

Physics-aware Difference Graph Networks for Sparsely-Observed Dynamics

- Sungyong Seo, *Chuizheng Meng*, Yan Liu
- abstract@open-review(Poster): Sparsely available data points cause numerical error on finite differences which hinders us from modeling the dynamics of physical systems. The discretization error becomes even larger when the sparse data are irregularly distributed or defined on an unstructured grid, making it hard to build deep learning models to handle physics-governing observations on the unstructured grid. In this paper, we propose a novel architecture, Physics-aware Difference Graph Networks (PA-DGN), which exploits neighboring information to learn finite differences inspired by physics equations. PA-DGN leverages data-driven end-to-end learning to discover underlying dynamical relations between the spatial and temporal differences in given sequential observations. We demonstrate the superiority of PA-DGN in the approximation of directional derivatives and the prediction of graph signals on the synthetic data and the real-world climate observations from weather stations.

Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks

- Ziwei Ji, Matus Telgarsky
- abstract@open-review(Poster): Recent theoretical work has guaranteed that overparameterized networks trained by gradient descent achieve arbitrarily low training error, and sometimes even low test error. The required width, however, is always polynomial in at least one of the sample size n, the (inverse) target error $1/\epsilon$, and the (inverse) failure probability $1/\delta$. This work shows that $\widetilde{\Theta}(1/\epsilon)$ iterations of gradient descent with $\widetilde{\Omega}(1/\epsilon^2)$ training examples on two-layer ReLU networks of any width exceeding $\textrm{polylog}(n,1/\epsilon,1/\delta)$ suffice to achieve a test misclassification error of ϵ. We also prove that stochastic gradient descent can achieve ϵ test error with polylogarithmic width and $\widetilde{\Theta}(1/\epsilon)$ samples. The analysis relies upon the separation margin of the limiting kernel, which is guaranteed positive, can distinguish between true labels and random labels, and can give a tight sample-complexity analysis in the infinite-width setting.

Learning representations for binary-classification without backpropagation

- Mathias Lechner
- abstract@open-review(Poster): The family of feedback alignment (FA) algorithms aims to provide a more biologically motivated alternative to backpropagation (BP), by substituting the computations that are unrealistic to be implemented in physical brains. While FA algorithms have been shown to work well in practice, there is a lack of rigorous theory proofing their learning capabilities.
  Here we introduce the first feedback alignment algorithm with provable learning guarantees. In contrast to existing work, we do not require any assumption about the size or depth of the network except that it has a single output neuron, i.e., such as for binary classification tasks. We show that our FA algorithm can deliver its theoretical promises in practice, surpassing the learning performance of existing FA methods and matching backpropagation in binary classification tasks. Finally, we demonstrate the limits of our FA variant when the number of output neurons grows beyond a certain quantity.

Mathematical Reasoning in Latent Space

- Dennis Lee, Christian Szegedy, Markus Rabe, Sarah Loos, Kshitij Bansal
- abstract@open-review(Talk): We design and conduct a simple experiment to study whether neural networks can perform several steps of approximate reasoning in a fixed dimensional latent space. The set of rewrites (i.e. transformations) that can be successfully performed on a statement represents essential semantic features of the statement. We can compress this information by embedding the formula in a vector space, such that the vector associated

with a statement can be used to predict whether a statement can be rewritten by other theorems. Predicting the embedding of a formula generated by some rewrite rule is naturally viewed as approximate reasoning in the latent space. In order to measure the effectiveness of this reasoning, we perform approximate deduction sequences in the latent space and use the resulting embedding to inform the semantic features of the corresponding formal statement (which is obtained by performing the corresponding rewrite sequence using real formulas). Our experiments show that graph neural networks can make non-trivial predictions about the rewrite-success of statements, even when they propagate predicted latent representations for several steps. Since our corpus of mathematical formulas includes a wide variety of mathematical disciplines, this experiment is a strong indicator for the feasibility of deduction in latent space in general.

Frequency-based Search-control in Dyna

- Yangchen Pan, Jincheng Mei, Amir-massoud Farahmand
- abstract@open-review(Poster): Model-based reinforcement learning has been empirically demonstrated as a successful strategy to improve sample efficiency. In particular, Dyna is an elegant model-based architecture integrating learning and planning that provides huge flexibility of using a model. One of the most important components in Dyna is called search-control, which refers to the process of generating state or state-action pairs from which we query the model to acquire simulated experiences. Search-control is critical in improving learning efficiency. In this work, we propose a simple and novel search-control strategy by searching high frequency regions of the value function. Our main intuition is built on Shannon sampling theorem from signal processing, which indicates that a high frequency signal requires more samples to reconstruct. We empirically show that a high frequency function is more difficult to approximate. This suggests a search-control strategy: we should use states from high frequency regions of the value function to query the model to acquire more samples. We develop a simple strategy to locally measure the frequency of a function by gradient and hessian norms, and provide theoretical justification for this approach. We then apply our strategy to search-control in Dyna, and conduct experiments to show its property and effectiveness on benchmark domains.

Towards Stable and Efficient Training of Verifiably Robust Neural Networks

- Huan Zhang, Hongge Chen, Chaowei Xiao, Sven Gowal, Robert Stanforth, Bo Li, Duane Boning, Cho-Jui Hsieh
- abstract@open-review(Poster): Training neural networks with verifiable robustness guarantees is challenging. Several existing approaches utilize linear relaxation based neural network output bounds under perturbation, but they can slow down training by a factor of hundreds depending on the underlying network architectures. Meanwhile, interval bound propagation (IBP) based training is efficient and significantly outperforms linear relaxation based methods on many tasks, yet it may suffer from stability issues since the bounds are much looser especially at the beginning of training. In this paper, we propose a new certified adversarial training method, CROWN-IBP, by combining the fast IBP bounds in a forward bounding pass and a tight linear relaxation based bound, CROWN, in a backward bounding pass. CROWN-IBP is computationally efficient and consistently outperforms IBP baselines on training verifiably robust neural networks. We conduct large scale experiments on MNIST and CIFAR datasets, and outperform all previous linear relaxation and bound propagation based certified defenses in L_inf robustness. Notably, we achieve 7.02% verified test error on MNIST at epsilon=0.3, and 66.94% on CIFAR-10 with epsilon=8/255.

Iterative energy-based projection on a normal data manifold for anomaly localization

- David Dehaene, Oriel Frigo, Sébastien Combrexelle, Pierre Eline
- abstract@open-review(Poster): Autoencoder reconstructions are widely used for the task of unsupervised anomaly localization. Indeed, an autoencoder trained on normal data is expected to only be able to reconstruct normal features of the data, allowing the segmentation of anomalous pixels in an image via a simple comparison between the image and its autoencoder reconstruction. In practice however, local defects added to a normal image can deteriorate the whole reconstruction, making this segmentation challenging. To tackle the issue, we propose in this paper a new approach for projecting anomalous data on a autoencoder-learned normal data manifold, by using gradient descent on an energy derived from the autoencoder's loss function. This energy can be augmented with regularization terms that model priors on what constitutes the user-defined optimal projection. By iteratively updating the input of the autoencoder, we bypass the loss of high-frequency information caused by the autoencoder bottleneck. This allows to produce images of higher quality than classic reconstructions. Our method achieves state-of-the-art results on various anomaly localization datasets. It also shows promising results at an inpainting task on the CelebA dataset.

Towards neural networks that provably know when they don't know

- Alexander Meinke, Matthias Hein
- abstract@open-review(Poster): It has recently been shown that ReLU networks produce arbitrarily over-confident predictions far away from the training data. Thus, ReLU networks do not know when they don't know. However, this is a highly important property in safety critical applications. In the context of out-of-distribution detection (OOD) there have been a number of proposals to mitigate this problem but none of them are able to make any mathematical guarantees. In this paper we propose a new approach to OOD which overcomes both problems. Our approach can be used with ReLU networks and provides provably low confidence predictions far away from the training data as well as the first certificates for low confidence predictions in a neighborhood of an out-distribution point. In the experiments we show that state-of-the-art methods fail in this worst-case setting whereas our model can guarantee its performance while retaining state-of-the-art OOD performance.

BatchEnsemble: an Alternative Approach to Efficient Ensemble and Lifelong Learning

- Yeming Wen, Dustin Tran, Jimmy Ba
- abstract@open-review(Poster): Ensembles, where multiple neural networks are trained individually and their predictions are averaged, have been shown to be widely successful for improving both the accuracy and predictive uncertainty of single neural networks. However, an ensemble's cost for both training and testing increases linearly with the number of networks, which quickly becomes untenable. In this paper, we propose BatchEnsemble, an ensemble method whose computational and memory costs are significantly lower than typical ensembles. BatchEnsemble achieves this by defining each weight matrix to be the Hadamard product of a shared weight among all ensemble members and a rank-one matrix per member. Unlike ensembles, BatchEnsemble is not only parallelizable across devices, where one device trains one member, but also parallelizable within a device, where multiple ensemble members are updated simultaneously for a given mini-batch. Across CIFAR-10, CIFAR-100, WMT14 EN-DE/EN-FR translation, and out-of-distribution tasks, BatchEnsemble yields competitive accuracy and uncertainties as typical ensembles; the speedup at test time is 3X and memory reduction is 3X at an ensemble of size 4. We also apply BatchEnsemble to lifelong learning, where on Split-CIFAR-100, BatchEnsemble yields comparable performance to progressive neural networks while having a much lower computational and memory costs. We further show that BatchEnsemble can easily scale up to lifelong learning on Split-ImageNet which involves 100 sequential learning tasks

Inductive representation learning on temporal graphs

- da Xu, chuanwei ruan, evren korpeoglu, sushant kumar, kannan achan
- abstract@open-review(Poster): Inductive representation learning on temporal graphs is an important step toward salable machine learning on real-world dynamic networks. The evolving nature of temporal dynamic graphs requires handling new nodes as well as capturing temporal patterns. The node embeddings, which are now functions of time, should represent both the static node features and the evolving topological structures. Moreover, node and topological features can be temporal as well, whose patterns the node embeddings should also capture. We propose the temporal graph attention (TGAT) layer to efficiently aggregate temporal-topological neighborhood features to learn the time-feature interactions. For TGAT, we use the self-attention

mechanism as building block and develop a novel functional time encoding technique based on the classical Bochner's theorem from harmonic analysis. By stacking TGAT layers, the network recognizes the node embeddings as functions of time and is able to inductively infer embeddings for both new and observed nodes as the graph evolves. The proposed approach handles both node classification and link prediction task, and can be naturally extended to include the temporal edge features. We evaluate our method with transductive and inductive tasks under temporal settings with two benchmark and one industrial dataset. Our TGAT model compares favorably to state-of-the-art baselines as well as the previous temporal graph embedding approaches.

A Probabilistic Formulation of Unsupervised Text Style Transfer

- Junxian He, Xinyi Wang, Graham Neubig, Taylor Berg-Kirkpatrick
- abstract@open-review(Spotlight): We present a deep generative model for unsupervised text style transfer that unifies previously proposed non-generative techniques. Our probabilistic approach models non-parallel data from two domains as a partially observed parallel corpus. By hypothesizing a parallel latent sequence that generates each observed sequence, our model learns to transform sequences from one domain to another in a completely unsupervised fashion. In contrast with traditional generative sequence models (e.g. the HMM), our model makes few assumptions about the data it generates: it uses a recurrent language model as a prior and an encoder-decoder as a transduction distribution. While computation of marginal data likelihood is intractable in this model class, we show that amortized variational inference admits a practical surrogate. Further, by drawing connections between our variational objective and other recent unsupervised style transfer and machine translation techniques, we show how our probabilistic view can unify some known non-generative objectives such as backtranslation and adversarial loss. Finally, we demonstrate the effectiveness of our method on a wide range of unsupervised style transfer tasks, including sentiment transfer, formality transfer, word decipherment, author imitation, and related language translation. Across all style transfer tasks, our approach yields substantial gains over state-of-the-art non-generative baselines, including the state-of-the-art unsupervised machine translation techniques that our approach generalizes. Further, we conduct experiments on a standard unsupervised machine translation task and find that our unified approach matches the current state-of-the-art.

Generative Models for Effective ML on Private, Decentralized Datasets

- Sean Augenstein, H. Brendan McMahan, Daniel Ramage, Swaroop Ramaswamy, Peter Kairouz, Mingqing Chen, Rajiv Mathews, Blaise Aguera y Arcas
- abstract@open-review(Poster): To improve real-world applications of machine learning, experienced modelers develop intuition about their datasets, their models, and how the two interact. Manual inspection of raw data—of representative samples, of outliers, of misclassifications—is an essential tool in a) identifying and fixing problems in the data, b) generating new modeling hypotheses, and c) assigning or refining human-provided labels. However, manual data inspection is risky for privacy-sensitive datasets, such as those representing the behavior of real-world individuals. Furthermore, manual data inspection is impossible in the increasingly important setting of federated learning, where raw examples are stored at the edge and the modeler may only access aggregated outputs such as metrics or model parameters. This paper demonstrates that generative models—trained using federated methods and with formal differential privacy guarantees—can be used effectively to debug data issues even when the data cannot be directly inspected. We explore these methods in applications to text with differentially private federated RNNs and to images using a novel algorithm for differentially private federated GANs.

Picking Winning Tickets Before Training by Preserving Gradient Flow

- Chaoqi Wang, Guodong Zhang, Roger Grosse
- abstract@open-review(Poster): Overparameterization has been shown to benefit both the optimization and generalization of neural networks, but large networks are resource hungry at both training and test time. Network pruning can reduce test-time resource requirements, but is typically applied to trained networks and therefore cannot avoid the expensive training process. We aim to prune networks at initialization, thereby saving resources at training time as well. Specifically, we argue that efficient training requires preserving the gradient flow through the network. This leads to a simple but effective pruning criterion we term Gradient Signal Preservation (GraSP). We empirically investigate the effectiveness of the proposed method with extensive experiments on CIFAR-10, CIFAR-100, Tiny-ImageNet and ImageNet, using VGGNet and ResNet architectures. Our method can prune 80% of the weights of a VGG-16 network on ImageNet at initialization, with only a 1.6% drop in top-1 accuracy. Moreover, our method achieves significantly better performance than the baseline at extreme sparsity levels. Our code is made public at: https://github.com/alecwangcq/GraSP.

Curriculum Loss: Robust Learning and Generalization against Label Corruption

- Yueming Lyu, Ivor W. Tsang
- abstract@open-review(Poster): Deep neural networks (DNNs) have great expressive power, which can even memorize samples with wrong labels. It is vitally important to reiterate robustness and generalization in DNNs against label corruption. To this end, this paper studies the 0-1 loss, which has a monotonic relationship between empirical adversary (reweighted) risk (Hu et al. 2018). Although the 0-1 loss is robust to outliers, it is also difficult to optimize. To efficiently optimize the 0-1 loss while keeping its robust properties, we propose a very simple and efficient loss, i.e. curriculum loss (CL). Our CL is a tighter upper bound of the 0-1 loss compared with conventional summation based surrogate losses. Moreover, CL can adaptively select samples for stagewise training. As a result, our loss can be deemed as a novel perspective of curriculum sample selection strategy, which bridges a connection between curriculum learning and robust learning. Experimental results on noisy MNIST, CIFAR10 and CIFAR100 dataset validate the robustness of the proposed loss.

Uncertainty-guided Continual Learning with Bayesian Neural Networks

- Sayna Ebrahimi, Mohamed Elhoseiny, Trevor Darrell, Marcus Rohrbach
- abstract@open-review(Poster): Continual learning aims to learn new tasks without forgetting previously learned ones. This is especially challenging when one cannot access data from previous tasks and when the model has a fixed capacity. Current regularization-based continual learning algorithms need an external representation and extra computation to measure the parameters' \textit{importance}. In contrast, we propose Uncertainty-guided Continual Bayesian Neural Networks (UCB), where the learning rate adapts according to the uncertainty defined in the probability distribution of the weights in networks. Uncertainty is a natural way to identify \textit{what to remember} and \textit{what to change} as we continually learn, and thus mitigate catastrophic forgetting. We also show a variant of our model, which uses uncertainty for weight pruning and retains task performance after pruning by saving binary masks per tasks. We evaluate our UCB approach extensively on diverse object classification datasets with short and long sequences of tasks and report superior or on-par performance compared to existing approaches. Additionally, we show that our model does not necessarily need task information at test time, i.e. it does not presume knowledge of which task a sample belongs to.

Training Recurrent Neural Networks Online by Learning Explicit State Variables

- Somjit Nath, Vincent Liu, Alan Chan, Xin Li, Adam White, Martha White
- abstract@open-review(Poster): Recurrent neural networks (RNNs) allow an agent to construct a state-representation from a stream of experience, which is essential in partially observable problems. However, there are two primary issues one must overcome when training an RNN: the sensitivity of the learning algorithm's performance to truncation length and and long training times. There are variety of strategies to improve training in RNNs, the mostly notably Backprop Through Time (BPTT) and by Real-Time Recurrent Learning. These strategies, however, are typically computationally expensive and focus computation on computing gradients back in time. In this work, we reformulate the RNN training objective to explicitly learn state vectors; this breaks the dependence across time and so avoids the need to estimate gradients far back in time. We show that for a fixed buffer of data, our algorithm---

called Fixed Point Propagation (FPP)---is sound: it converges to a stationary point of the new objective. We investigate the empirical performance of our online FPP algorithm, particularly in terms of computation compared to truncated BPTT with varying truncation levels.

Cross-lingual Alignment vs Joint Training: A Comparative Study and A Simple Unified Framework

- Zirui Wang, *Jiateng Xie*, Ruochen Xu, Yiming Yang, Graham Neubig, Jaime G. Carbonell
- abstract@open-review(Poster): Learning multilingual representations of text has proven a successful method for many cross-lingual transfer learning tasks. There are two main paradigms for learning such representations: (1) alignment, which maps different independently trained monolingual representations into a shared space, and (2) joint training, which directly learns unified multilingual representations using monolingual and cross-lingual objectives jointly. In this paper, we first conduct direct comparisons of representations learned using both of these methods across diverse cross-lingual tasks. Our empirical results reveal a set of pros and cons for both methods, and show that the relative performance of alignment versus joint training is task-dependent. Stemming from this analysis, we propose a simple and novel framework that combines these two previously mutually-exclusive approaches. Extensive experiments demonstrate that our proposed framework alleviates limitations of both approaches, and outperforms existing methods on the MUSE bilingual lexicon induction (BLI) benchmark. We further show that this framework can generalize to contextualized representations such as Multilingual BERT, and produces state-of-the-art results on the CoNLL cross-lingual NER benchmark.

Robust Reinforcement Learning for Continuous Control with Model Misspecification

- Daniel J. Mankowitz, Nir Levine, Rae Jeong, Abbas Abdolmaleki, Jost Tobias Springenberg, Yuanyuan Shi, Jackie Kay, Todd Hester, Timothy Mann, Martin Riedmiller
- abstract@open-review(Poster): We provide a framework for incorporating robustness -- to perturbations in the transition dynamics which we refer to as model misspecification -- into continuous control Reinforcement Learning (RL) algorithms. We specifically focus on incorporating robustness into a state-of-the-art continuous control RL algorithm called Maximum a-posteriori Policy Optimization (MPO). We achieve this by learning a policy that optimizes for a worst case, entropy-regularized, expected return objective and derive a corresponding robust entropy-regularized Bellman contraction operator. In addition, we introduce a less conservative, soft-robust, entropy-regularized objective with a corresponding Bellman operator. We show that both, robust and soft-robust policies, outperform their non-robust counterparts in nine Mujoco domains with environment perturbations. In addition, we show improved robust performance on a challenging, simulated, dexterous robotic hand. Finally, we present multiple investigative experiments that provide a deeper insight into the robustness framework; including an adaptation to another continuous control RL algorithm. Performance videos can be found online at https://sites.google.com/view/robust-rl.

Decoupling Representation and Classifier for Long-Tailed Recognition

- Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, Yannis Kalantidis
- abstract@open-review(Poster): The long-tail distribution of the visual world poses great challenges for deep learning based classification models on how to handle the class imbalance problem. Existing solutions usually involve class-balancing strategies, e.g., by loss re-weighting, data re-sampling, or transfer learning from head- to tail-classes, but most of them adhere to the scheme of jointly learning representations and classifiers. In this work, we decouple the learning procedure into representation learning and classification, and systematically explore how different balancing strategies affect them for long-tailed recognition. The findings are surprising: (1) data imbalance might not be an issue in learning high-quality representations; (2) with representations learned with the simplest instance-balanced (natural) sampling, it is also possible to achieve strong long-tailed recognition ability by adjusting only the classifier. We conduct extensive experiments and set new state-of-the-art performance on common long-tailed benchmarks like ImageNet-LT, Places-LT and iNaturalist, showing that it is possible to outperform carefully designed losses, sampling strategies, even complex modules with memory, by using a straightforward approach that decouples representation and classification. Our code is available at https://github.com/facebookresearch/classifier-balancing.

Learning from Unlabelled Videos Using Contrastive Predictive Neural 3D Mapping

- Adam W. Harley, Shrinidhi K. Lakshmikanth, Fangyu Li, Xian Zhou, Hsiao-Yu Fish Tung, Katerina Fragkiadaki
- abstract@open-review(Poster): Predictive coding theories suggest that the brain learns by predicting observations at various levels of abstraction. One of the most basic prediction tasks is view prediction: how would a given scene look from an alternative viewpoint? Humans excel at this task. Our ability to imagine and fill in missing information is tightly coupled with perception: we feel as if we see the world in 3 dimensions, while in fact, information from only the front surface of the world hits our retinas. This paper explores the role of view prediction in the development of 3D visual recognition. We propose neural 3D mapping networks, which take as input 2.5D (color and depth) video streams captured by a moving camera, and lift them to stable 3D feature maps of the scene, by disentangling the scene content from the motion of the camera. The model also projects its 3D feature maps to novel viewpoints, to predict and match against target views. We propose contrastive prediction losses to replace the standard color regression loss, and show that this leads to better performance on complex photorealistic data. We show that the proposed model learns visual representations useful for (1) semi-supervised learning of 3D object detectors, and (2) unsupervised learning of 3D moving object detectors, by estimating the motion of the inferred 3D feature maps in videos of dynamic scenes. To the best of our knowledge, this is the first work that empirically shows view prediction to be a scalable self-supervised task beneficial to 3D object detection.

Dream to Control: Learning Behaviors by Latent Imagination

- Danijar Hafner, Timothy Lillicrap, Jimmy Ba, Mohammad Norouzi
- abstract@open-review(Spotlight): Learned world models summarize an agent's experience to facilitate learning complex behaviors. While learning world models from high-dimensional sensory inputs is becoming feasible through deep learning, there are many potential ways for deriving behaviors from them. We present Dreamer, a reinforcement learning agent that solves long-horizon tasks from images purely by latent imagination. We efficiently learn behaviors by propagating analytic gradients of learned state values back through trajectories imagined in the compact state space of a learned world model. On 20 challenging visual control tasks, Dreamer exceeds existing approaches in data-efficiency, computation time, and final performance.

From Inference to Generation: End-to-end Fully Self-supervised Generation of Human Face from Speech

- Hyeong-Seok Choi, Changdae Park, Kyogu Lee
- abstract@open-review(Poster): This work seeks the possibility of generating the human face from voice solely based on the audio-visual data without any human-labeled annotations. To this end, we propose a multi-modal learning framework that links the inference stage and generation stage. First, the inference networks are trained to match the speaker identity between the two different modalities. Then the pre-trained inference networks cooperate with the generation network by giving conditional information about the voice. The proposed method exploits the recent development of GANs techniques and generates the human face directly from the speech waveform making our system fully end-to-end. We analyze the extent to which the network can naturally disentangle two latent factors that contribute to the generation of a face image one that comes directly from a speech signal and the other that is not related to it and explore whether the network can learn to generate natural human face image distribution by modeling these factors. Experimental results show that the proposed network can not only match the relationship between the human face and speech, but can also generate the high-quality human face sample conditioned on its speech. Finally, the correlation between the generated face and the corresponding speech is quantitatively measured to analyze the relationship between the two modalities.

Real or Not Real, that is the Question

- Yuanbo Xiangli, *Yubin Deng*, Bo Dai*, Chen Change Loy, Dahua Lin
- abstract@open-review(Spotlight): While generative adversarial networks (GAN) have been widely adopted in various topics, in this paper we generalize the standard GAN to a new perspective by treating realness as a random variable that can be estimated from multiple angles. In this generalized framework, referred to as RealnessGAN, the discriminator outputs a distribution as the measure of realness. While RealnessGAN shares similar theoretical guarantees with the standard GAN, it provides more insights on adversarial learning. More importantly, compared to multiple baselines, RealnessGAN provides stronger guidance for the generator, achieving improvements on both synthetic and real-world datasets. Moreover, it enables the basic DCGAN architecture to generate realistic images at 1024*1024 resolution when trained from scratch.

LambdaNet: Probabilistic Type Inference using Graph Neural Networks

- Jiayi Wei, Maruth Goyal, Greg Durrett, Isil Dillig
- abstract@open-review(Poster): As gradual typing becomes increasingly popular in languages like Python and TypeScript, there is a growing need to infer type annotations automatically. While type annotations help with tasks like code completion and static error catching, these annotations cannot be fully inferred by compilers and are tedious to annotate by hand. This paper proposes a probabilistic type inference scheme for TypeScript based on a graph neural network. Our approach first uses lightweight source code analysis to generate a program abstraction called a type dependency graph, which links type variables with logical constraints as well as name and usage information. Given this program abstraction, we then use a graph neural network to propagate information between related type variables and eventually make type predictions. Our neural architecture can predict both standard types, like number or string, as well as user-defined types that have not been encountered during training. Our experimental results show that our approach outperforms prior work in this space by 14% (absolute) on library types, while having the ability to make type predictions that are out of scope for existing techniques.

Model-Augmented Actor-Critic: Backpropagating through Paths

- Ignasi Clavera, Yao Fu, Pieter Abbeel
- abstract@open-review(Poster): Current model-based reinforcement learning approaches use the model simply as a learned black-box simulator to augment the data for policy optimization or value function learning. In this paper, we show how to make more effective use of the model by exploiting its differentiability. We construct a policy optimization algorithm that uses the pathwise derivative of the learned model and policy across future timesteps. Instabilities of learning across many timesteps are prevented by using a terminal value function, learning the policy in an actor-critic fashion. Furthermore, we present a derivation on the monotonic improvement of our objective in terms of the gradient error in the model and value function. We show that our approach (i) is consistently more sample efficient than existing state-of-the-art model-based algorithms, (ii) matches the asymptotic performance of model-free algorithms, and (iii) scales to long horizons, a regime where typically past model-based approaches have struggled.

Neural Symbolic Reader: Scalable Integration of Distributed and Symbolic Representations for Reading Comprehension

- Xinyun Chen, Chen Liang, Adams Wei Yu, Denny Zhou, Dawn Song, Quoc V. Le
- abstract@open-review(Spotlight): Integrating distributed representations with symbolic operations is essential for reading comprehension requiring complex reasoning, such as counting, sorting and arithmetics, but most existing approaches are hard to scale to more domains or more complex reasoning. In this work, we propose the Neural Symbolic Reader (NeRd), which includes a reader, e.g., BERT, to encode the passage and question, and a programmer, e.g., LSTM, to generate a program that is executed to produce the answer. Compared to previous works, NeRd is more scalable in two aspects: (1) domain-agnostic, i.e., the same neural architecture works for different domains; (2) compositional, i.e., when needed, complex programs can be generated by recursively applying the predefined operators, which become executable and interpretable representations for more complex reasoning. Furthermore, to overcome the challenge of training NeRd with weak supervision, we apply data augmentation techniques and hard Expectation-Maximization (EM) with thresholding. On DROP, a challenging reading comprehension dataset that requires discrete reasoning, NeRd achieves 1.37%/1.18% absolute improvement over the state-of-the-art on EM/F1 metrics. With the same architecture, NeRd significantly outperforms the baselines on MathQA, a math problem benchmark that requires multiple steps of reasoning, by 25.5% absolute increment on accuracy when trained on all the annotated programs. More importantly, NeRd still beats the baselines even when only 20% of the program annotations are given.

Variational Autoencoders for Highly Multivariate Spatial Point Processes Intensities

- Baichuan Yuan, Xiaowei Wang, Jianxin Ma, Chang Zhou, Andrea L. Bertozzi, Hongxia Yang
- abstract@open-review(Poster): Multivariate spatial point process models can describe heterotopic data over space. However, highly multivariate intensities are computationally challenging due to the curse of dimensionality. To bridge this gap, we introduce a declustering based hidden variable model that leads to an efficient inference procedure via a variational autoencoder (VAE). We also prove that this model is a generalization of the VAE-based model for collaborative filtering. This leads to an interesting application of spatial point process models to recommender systems. Experimental results show the method's utility on both synthetic data and real-world data sets.

Denoising and Regularization via Exploiting the Structural Bias of Convolutional Generators

- Reinhard Heckel and Mahdi Soltanolkotabi
- abstract@open-review(Poster): Convolutional Neural Networks (CNNs) have emerged as highly successful tools for image generation, recovery, and restoration. A major contributing factor to this success is that convolutional networks impose strong prior assumptions about natural images. A surprising experiment that highlights this architectural bias towards natural images is that one can remove noise and corruptions from a natural image without using any training data, by simply fitting (via gradient descent) a randomly initialized, over-parameterized convolutional generator to the corrupted image. While this over-parameterized network can fit the corrupted image perfectly, surprisingly after a few iterations of gradient descent it generates an almost uncorrupted image. This intriguing phenomenon enables state-of-the-art CNN-based denoising and regularization of other inverse problems. In this paper, we attribute this effect to a particular architectural choice of convolutional networks, namely convolutions with fixed interpolating filters. We then formally characterize the dynamics of fitting a two-layer convolutional generator to a noisy signal and prove that early-stopped gradient descent denoises/regularizes. Our proof relies on showing that convolutional generators fit the structured part of an image significantly faster than the corrupted portion.

Network Deconvolution

- Chengxi Ye, Matthew Evanusa, Hua He, Anton Mitrokhin, Tom Goldstein, James A. Yorke, Cornelia Fermuller, Yiannis Aloimonos
- abstract@open-review(Spotlight): Convolution is a central operation in Convolutional Neural Networks (CNNs), which applies a kernel to overlapping regions shifted across the image. However, because of the strong correlations in real-world image data, convolutional kernels are in effect re-learning redundant data. In this work, we show that this redundancy has made neural network training challenging, and propose network deconvolution, a procedure which optimally removes pixel-wise and channel-wise correlations before the data is fed into each layer. Network deconvolution can be efficiently calculated at a fraction of the computational cost of a convolution layer. We also show that the deconvolution filters in the first layer of the network resemble the center-surround structure found in biological neurons in the visual regions of the brain. Filtering with such kernels results in a sparse

representation, a desired property that has been missing in the training of neural networks. Learning from the sparse representation promotes faster convergence and superior results without the use of batch normalization. We apply our network deconvolution operation to 10 modern neural network models by replacing batch normalization within each. Extensive experiments show that the network deconvolution operation is able to deliver performance improvement in all cases on the CIFAR-10, CIFAR-100, MNIST, Fashion-MNIST, Cityscapes, and ImageNet datasets.

Revisiting Self-Training for Neural Sequence Generation

- Junxian He, Jiatao Gu, Jiajun Shen, Marc'Aurelio Ranzato
- abstract@open-review(Poster): Self-training is one of the earliest and simplest semi-supervised methods. The key idea is to augment the original labeled dataset with unlabeled data paired with the model's prediction (i.e. the pseudo-parallel data). While self-training has been extensively studied on classification problems, in complex sequence generation tasks (e.g. machine translation) it is still unclear how self-training works due to the compositionality of the target space. In this work, we first empirically show that self-training is able to decently improve the supervised baseline on neural sequence generation tasks. Through careful examination of the performance gains, we find that the perturbation on the hidden states (i.e. dropout) is critical for self-training to benefit from the pseudo-parallel data, which acts as a regularizer and forces the model to yield close predictions for similar unlabeled inputs. Such effect helps the model correct some incorrect predictions on unlabeled data. To further encourage this mechanism, we propose to inject noise to the input space, resulting in a noisy version of self-training. Empirical study on standard machine translation and text summarization benchmarks shows that noisy self-training is able to effectively utilize unlabeled data and improve the performance of the supervised baseline by a large margin.

Meta-Q-Learning

- Rasool Fakoor, Pratik Chaudhari, Stefano Soatto, Alexander J. Smola
- abstract@open-review(Talk): This paper introduces Meta-Q-Learning (MQL), a new off-policy algorithm for meta-Reinforcement Learning (meta-RL). MQL builds upon three simple ideas. First, we show that Q-learning is competitive with state-of-the-art meta-RL algorithms if given access to a context variable that is a representation of the past trajectory. Second, a multi-task objective to maximize the average reward across the training tasks is an effective method to meta-train RL policies. Third, past data from the meta-training replay buffer can be recycled to adapt the policy on a new task using off-policy updates. MQL draws upon ideas in propensity estimation to do so and thereby amplifies the amount of available data for adaptation. Experiments on standard continuous-control benchmarks suggest that MQL compares favorably with the state of the art in meta-RL.

Towards a Deep Network Architecture for Structured Smoothness

- Haroun Habeeb, Oluwasanmi Koyejo
- abstract@open-review(Poster): We propose the Fixed Grouping Layer (FGL); a novel feedforward layer designed to incorporate the inductive bias of structured smoothness into a deep learning model. FGL achieves this goal by connecting nodes across layers based on spatial similarity. The use of structured smoothness, as implemented by FGL, is motivated by applications to structured spatial data, which is, in turn, motivated by domain knowledge. The proposed model architecture outperforms conventional neural network architectures across a variety of simulated and real datasets with structured smoothness.

On the Global Convergence of Training Deep Linear ResNets

- Difan Zou, Philip M. Long, Quanquan Gu
- abstract@open-review(Poster): We study the convergence of gradient descent (GD) and stochastic gradient descent (SGD) for training L-hidden-layer linear residual networks (ResNets). We prove that for training deep residual networks with certain linear transformations at input and output layers, which are fixed throughout training, both GD and SGD with zero initialization on all hidden weights can converge to the global minimum of the training loss. Moreover, when specializing to appropriate Gaussian random linear transformations, GD and SGD provably optimize wide enough deep linear ResNets. Compared with the global convergence result of GD for training standard deep linear networks \citep{du2019width}, our condition on the neural network width is sharper by a factor of $O(\kappa L)$, where κ denotes the condition number of the covariance matrix of the training data. We further propose a modified identity input and output transformations, and show that a $(d+k)$-wide neural network is sufficient to guarantee the global convergence of GD/SGD, where d,k are the input and output dimensions respectively.

A Closer Look at the Optimization Landscapes of Generative Adversarial Networks

- Hugo Berard, Gauthier Gidel, Amjad Almahairi, Pascal Vincent, Simon Lacoste-Julien
- abstract@open-review(Poster): Generative adversarial networks have been very successful in generative modeling, however they remain relatively challenging to train compared to standard deep neural networks. In this paper, we propose new visualization techniques for the optimization landscapes of GANs that enable us to study the game vector field resulting from the concatenation of the gradient of both players. Using these visualization techniques we try to bridge the gap between theory and practice by showing empirically that the training of GANs exhibits significant rotations around LSSP, similar to the one predicted by theory on toy examples. Moreover, we provide empirical evidence that GAN training seems to converge to a stable stationary point which is a saddle point for the generator loss, not a minimum, while still achieving excellent performance.

Simplified Action Decoder for Deep Multi-Agent Reinforcement Learning

- Hengyuan Hu, Jakob N Foerster
- abstract@open-review(Spotlight): In recent years we have seen fast progress on a number of benchmark problems in AI, with modern methods achieving near or super human performance in Go, Poker and Dota. One common aspect of all of these challenges is that they are by design adversarial or, technically speaking, zero-sum. In contrast to these settings, success in the real world commonly requires humans to collaborate and communicate with others, in settings that are, at least partially, cooperative. In the last year, the card game Hanabi has been established as a new benchmark environment for AI to fill this gap. In particular, Hanabi is interesting to humans since it is entirely focused on theory of mind, i.e. the ability to effectively reason over the intentions, beliefs and point of view of other agents when observing their actions. Learning to be informative when observed by others is an interesting challenge for Reinforcement Learning (RL): Fundamentally, RL requires agents to explore in order to discover good policies. However, when done naively, this randomness will inherently make their actions less informative to others during training. We present a new deep multi-agent RL method, the Simplified Action Decoder (SAD), which resolves this contradiction exploiting the centralized training phase. During training SAD allows other agents to not only observe the (exploratory) action chosen, but agents instead also observe the greedy action of their team mates. By combining this simple intuition with an auxiliary task for state prediction and best practices for multi-agent learning, SAD establishes a new state of the art for 2-5 players on the self-play part of the Hanabi challenge.

Biologically inspired sleep algorithm for increased generalization and adversarial robustness in deep neural networks

- Timothy Tadros, Giri Krishnan, Ramyaa Ramyaa, Maxim Bazhenov
- abstract@open-review(Poster): Current artificial neural networks (ANNs) can perform and excel at a variety of tasks ranging from image classification to spam detection through training on large datasets of labeled data. While the trained network may perform well on similar testing data, inputs that differ

even slightly from the training data may trigger unpredictable behavior. Due to this limitation, it is possible to design inputs with very small perturbations that can result in misclassification. These adversarial attacks present a security risk to deployed ANNs and indicate a divergence between how ANNs and humans perform classification. Humans are robust at behaving in the presence of noise and are capable of correctly classifying objects that are noisy, blurred, or otherwise distorted. It has been hypothesized that sleep promotes generalization of knowledge and improves robustness against noise in animals and humans. In this work, we utilize a biologically inspired sleep phase in ANNs and demonstrate the benefit of sleep on defending against adversarial attacks as well as in increasing ANN classification robustness. We compare the sleep algorithm's performance on various robustness tasks with two previously proposed adversarial defenses - defensive distillation and fine-tuning. We report an increase in robustness after sleep phase to adversarial attacks as well as to general image distortions for three datasets: MNIST, CUB200, and a toy dataset. Overall, these results demonstrate the potential for biologically inspired solutions to solve existing problems in ANNs and guide the development of more robust, human-like ANNs.

Distributed Bandit Learning: Near-Optimal Regret with Efficient Communication

- Yuanhao Wang, Jiachen Hu, Xiaoyu Chen, Liwei Wang
- abstract@open-review(Poster): We study the problem of regret minimization for distributed bandits learning, in which M agents work collaboratively to minimize their total regret under the coordination of a central server. Our goal is to design communication protocols with near-optimal regret and little communication cost, which is measured by the total amount of transmitted data. For distributed multi-armed bandits, we propose a protocol with near-optimal regret and only $O(M\log(MK))$ communication cost, where K is the number of arms. The communication cost is independent of the time horizon T, has only logarithmic dependence on the number of arms, and matches the lower bound except for a logarithmic factor. For distributed d-dimensional linear bandits, we propose a protocol that achieves near-optimal regret and has communication cost of order $O\left(\left(Md+d\log \log d\right)\log T\right)$, which has only logarithmic dependence on T.

Is a Good Representation Sufficient for Sample Efficient Reinforcement Learning?

- Simon S. Du, Sham M. Kakade, Ruosong Wang, Lin F. Yang
- abstract@open-review(Spotlight): Modern deep learning methods provide effective means to learn good representations. However, is a good representation itself sufficient for sample efficient reinforcement learning? This question has largely been studied only with respect to (worst-case) approximation error, in the more classical approximate dynamic programming literature. With regards to the statistical viewpoint, this question is largely unexplored, and the extant body of literature mainly focuses on conditions which \emph{permit} sample efficient reinforcement learning with little understanding of what are \emph{necessary} conditions for efficient reinforcement learning. This work shows that, from the statistical viewpoint, the situation is far subtler than suggested by the more traditional approximation viewpoint, where the requirements on the representation that suffice for sample efficient RL are even more stringent. Our main results provide sharp thresholds for reinforcement learning methods, showing that there are hard limitations on what constitutes good function approximation (in terms of the dimensionality of the representation), where we focus on natural representational conditions relevant to value-based, model-based, and policy-based learning. These lower bounds highlight that having a good (value-based, model-based, or policy-based) representation in and of itself is insufficient for efficient reinforcement learning, unless the quality of this approximation passes certain hard thresholds. Furthermore, our lower bounds also imply exponential separations on the sample complexity between 1) value-based learning with perfect representation and value-based learning with a good-but-not-perfect representation, 2) value-based learning and policy-based learning, 3) policy-based learning and supervised learning and 4) reinforcement learning and imitation learning.

Shifted and Squeezed 8-bit Floating Point format for Low-Precision Training of Deep Neural Networks

- Leopold Cambier, Anahita Bhiwandiwalla, Ting Gong, Oguz H. Elibol, Mehran Nekuii, Hanlin Tang
- abstract@open-review(Poster): Training with larger number of parameters while keeping fast iterations is an increasingly adopted strategy and trend for developing better performing Deep Neural Network (DNN) models. This necessitates increased memory footprint and computational requirements for training. Here we introduce a novel methodology for training deep neural networks using 8-bit floating point (FP8) numbers. Reduced bit precision allows for a larger effective memory and increased computational speed. We name this method Shifted and Squeezed FP8 (S2FP8). We show that, unlike previous 8-bit precision training methods, the proposed method works out of the box for representative models: ResNet50, Transformer and NCF. The method can maintain model accuracy without requiring fine-tuning loss scaling parameters or keeping certain layers in single precision. We introduce two learnable statistics of the DNN tensors - shifted and squeezed factors that are used to optimally adjust the range of the tensors in 8-bits, thus minimizing the loss in information due to quantization.

Intriguing Properties of Adversarial Training at Scale

- Cihang Xie, Alan Yuille
- abstract@open-review(Poster): Adversarial training is one of the main defenses against adversarial attacks. In this paper, we provide the first rigorous study on diagnosing elements of large-scale adversarial training on ImageNet, which reveals two intriguing properties.

First, we study the role of normalization. Batch normalization (BN) is a crucial element for achieving state-of-the-art performance on many vision tasks, but we show it may prevent networks from obtaining strong robustness in adversarial training. One unexpected observation is that, for models trained with BN, simply removing clean images from training data largely boosts adversarial robustness, i.e., 18.3%. We relate this phenomenon to the hypothesis that clean images and adversarial images are drawn from two different domains. This two-domain hypothesis may explain the issue of BN when training with a mixture of clean and adversarial images, as estimating normalization statistics of this mixture distribution is challenging. Guided by this two-domain hypothesis, we show disentangling the mixture distribution for normalization, i.e., applying separate BNs to clean and adversarial images for statistics estimation, achieves much stronger robustness. Additionally, we find that enforcing BNs to behave consistently at training and testing can further enhance robustness.

Second, we study the role of network capacity. We find our so-called "deep" networks are still shallow for the task of adversarial learning. Unlike traditional classification tasks where accuracy is only marginally improved by adding more layers to "deep" networks (e.g., ResNet-152), adversarial training exhibits a much stronger demand on deeper networks to achieve higher adversarial robustness. This robustness improvement can be observed substantially and consistently even by pushing the network capacity to an unprecedented scale, i.e., ResNet-638.

Deep Double Descent: Where Bigger Models and More Data Hurt

- Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, Ilya Sutskever
- abstract@open-review(Poster): We show that a variety of modern deep learning tasks exhibit a "double-descent" phenomenon where, as we increase model size, performance first gets worse and then gets better. Moreover, we show that double descent occurs not just as a function of model size, but also as a function of the number of training epochs. We unify the above phenomena by defining a new complexity measure we call the effective model complexity, and conjecture a generalized double descent with respect to this measure. Furthermore, our notion of model complexity allows us to identify certain regimes where increasing (even quadrupling) the number of train samples actually hurts test performance.

Decoding As Dynamic Programming For Recurrent Autoregressive Models

- Najam Zaidi, Trevor Cohn, Gholamreza Haffari

- abstract@open-review(Poster): Decoding in autoregressive models (ARMs) consists of searching for a high scoring output sequence under the trained model. Standard decoding methods, based on unidirectional greedy algorithm or beam search, are suboptimal due to error propagation and myopic decisions which do not account for future steps in the generation process. In this paper we present a novel decoding approach based on the method of auxiliary coordinates (Carreira-Perpinan & Wang, 2014) to address the aforementioned shortcomings. Our method introduces discrete variables for output tokens, and auxiliary continuous variables representing the states of the underlying ARM. The auxiliary variables lead to a factor graph approximation of the ARM, whose maximum a posteriori (MAP) inference is found exactly using dynamic programming. The MAP inference is then used to recreate an improved factor graph approximation of the ARM via updated auxiliary variables. We then extend our approach to decode in an ensemble of ARMs, possibly with different generation orders, which is out of reach for the standard unidirectional decoding algorithms. Experiments on the text infilling task over SWAG and Daily Dialogue datasets show that our decoding method is superior to strong unidirectional decoding baselines.

Synthesizing Programmatic Policies that Inductively Generalize

- Jeevana Priya Inala, Osbert Bastani, Zenna Tavares, Armando Solar-Lezama
- abstract@open-review(Poster): Deep reinforcement learning has successfully solved a number of challenging control tasks. However, learned policies typically have difficulty generalizing to novel environments. We propose an algorithm for learning programmatic state machine policies that can capture repeating behaviors. By doing so, they have the ability to generalize to instances requiring an arbitrary number of repetitions, a property we call inductive generalization. However, state machine policies are hard to learn since they consist of a combination of continuous and discrete structures. We propose a learning framework called adaptive teaching, which learns a state machine policy by imitating a teacher; in contrast to traditional imitation learning, our teacher adaptively updates itself based on the structure of the student. We show that our algorithm can be used to learn policies that inductively generalize to novel environments, whereas traditional neural network policies fail to do so.

Transformer-XH: Multi-Evidence Reasoning with eXtra Hop Attention

- Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul Bennett, Saurabh Tiwary
- abstract@open-review(Poster): Transformers have achieved new heights modeling natural language as a sequence of text tokens. However, in many real world scenarios, textual data inherently exhibits structures beyond a linear sequence such as trees and graphs; many tasks require reasoning with evidence scattered across multiple pieces of texts. This paper presents Transformer-XH, which uses eXtra Hop attention to enable intrinsic modeling of structured texts in a fully data-driven way. Its new attention mechanism naturally “hops” across the connected text sequences in addition to attending over tokens within each sequence. Thus, Transformer-XH better conducts joint multi-evidence reasoning by propagating information between documents and constructing global contextualized representations. On multi-hop question answering, Transformer-XH leads to a simpler multi-hop QA system which outperforms previous state-of-the-art on the HotpotQA FullWiki setting. On FEVER fact verification, applying Transformer-XH provides state-of-the-art accuracy and excels on claims whose verification requires multiple evidence.

Generalization through Memorization: Nearest Neighbor Language Models

- Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, Mike Lewis
- abstract@open-review(Poster): We introduce kNN-LMs, which extend a pre-trained neural language model (LM) by linearly interpolating it with a k-nearest neighbors (kNN) model. The nearest neighbors are computed according to distance in the pre-trained LM embedding space, and can be drawn from any text collection, including the original LM training data. Applying this transformation to a strong Wikitext-103 LM, with neighbors drawn from the original training set, our kNN-LM achieves a new state-of-the-art perplexity of 15.79 -- a 2.9 point improvement with no additional training. We also show that this approach has implications for efficiently scaling up to larger training sets and allows for effective domain adaptation, by simply varying the nearest neighbor datastore, again without further training. Qualitatively, the model is particularly helpful in predicting rare patterns, such as factual knowledge. Together, these results strongly suggest that learning similarity between sequences of text is easier than predicting the next word, and that nearest neighbor search is an effective approach for language modeling in the long tail.

Comparing Rewinding and Fine-tuning in Neural Network Pruning

- Alex Renda, Jonathan Frankle, Michael Carbin
- abstract@open-review(Talk): Many neural network pruning algorithms proceed in three steps: train the network to completion, remove unwanted structure to compress the network, and retrain the remaining structure to recover lost accuracy. The standard retraining technique, fine-tuning, trains the unpruned weights from their final trained values using a small fixed learning rate. In this paper, we compare fine-tuning to alternative retraining techniques. Weight rewinding (as proposed by Frankle et al., (2019)), rewinds unpruned weights to their values from earlier in training and retrains them from there using the original training schedule. Learning rate rewinding (which we propose) trains the unpruned weights from their final values using the same learning rate schedule as weight rewinding. Both rewinding techniques outperform fine-tuning, forming the basis of a network-agnostic pruning algorithm that matches the accuracy and compression ratios of several more network-specific state-of-the-art techniques.

Single Episode Policy Transfer in Reinforcement Learning

- Jiachen Yang, Brenden Petersen, Hongyuan Zha, Daniel Faissol
- abstract@open-review(Poster): Transfer and adaptation to new unknown environmental dynamics is a key challenge for reinforcement learning (RL). An even greater challenge is performing near-optimally in a single attempt at test time, possibly without access to dense rewards, which is not addressed by current methods that require multiple experience rollouts for adaptation. To achieve single episode transfer in a family of environments with related dynamics, we propose a general algorithm that optimizes a probe and an inference model to rapidly estimate underlying latent variables of test dynamics, which are then immediately used as input to a universal control policy. This modular approach enables integration of state-of-the-art algorithms for variational inference or RL. Moreover, our approach does not require access to rewards at test time, allowing it to perform in settings where existing adaptive approaches cannot. In diverse experimental domains with a single episode test constraint, our method significantly outperforms existing adaptive approaches and shows favorable performance against baselines for robust transfer.

Towards Stabilizing Batch Statistics in Backward Propagation of Batch Normalization

- Junjie Yan, Ruosi Wan, Xiangyu Zhang, Wei Zhang, Yichen Wei, Jian Sun
- abstract@open-review(Poster): Batch Normalization (BN) is one of the most widely used techniques in Deep Learning field. But its performance can awfully degrade with insufficient batch size. This weakness limits the usage of BN on many computer vision tasks like detection or segmentation, where batch size is usually small due to the constraint of memory consumption. Therefore many modified normalization techniques have been proposed, which either fail to restore the performance of BN completely, or have to introduce additional nonlinear operations in inference procedure and increase huge consumption. In this paper, we reveal that there are two extra batch statistics involved in backward propagation of BN, on which has never been well discussed before. The extra batch statistics associated with gradients also can severely affect the training of deep neural network. Based on our analysis, we propose a novel normalization method, named Moving Average Batch Normalization (MABN). MABN can completely restore the performance of vanilla BN in small batch cases, without introducing any additional nonlinear operations in inference procedure. We prove the benefits of MABN by both theoretical analysis and experiments. Our experiments demonstrate the effectiveness of MABN in multiple computer vision tasks including ImageNet and COCO. The code has been released in https://github.com/megvii-model/MABN.

NeurQuRI: Neural Question Requirement Inspector for Answerability Prediction in Machine Reading Comprehension

- Seohyun Back, Sai Chetan Chinthakindi, Akhil Kedia, Haejun Lee, Jaegul Choo
- abstract@open-review(Poster): Real-world question answering systems often retrieve potentially relevant documents to a given question through a keyword search, followed by a machine reading comprehension (MRC) step to find the exact answer from them. In this process, it is essential to properly determine whether an answer to the question exists in a given document. This task often becomes complicated when the question involves multiple different conditions or requirements which are to be met in the answer. For example, in a question "What was the projection of sea level increases in the fourth assessment report?", the answer should properly satisfy several conditions, such as "increases" (but not decreases) and "fourth" (but not third). To address this, we propose a neural question requirement inspection model called NeurQuRI that extracts a list of conditions from the question, each of which should be satisfied by the candidate answer generated by an MRC model. To check whether each condition is met, we propose a novel, attention-based loss function. We evaluate our approach on SQuAD 2.0 dataset by integrating the proposed module with various MRC models, demonstrating the consistent performance improvements across a wide range of state-of-the-art methods.

Learning The Difference That Makes A Difference With Counterfactually-Augmented Data

- Divyansh Kaushik, Eduard Hovy, Zachary Lipton
- abstract@open-review(Spotlight): Despite alarm over the reliance of machine learning systems on so-called spurious patterns, the term lacks coherent meaning in standard statistical frameworks. However, the language of causality offers clarity: spurious associations are due to confounding (e.g., a common cause), but not direct or indirect causal effects. In this paper, we focus on natural language processing, introducing methods and resources for training models less sensitive to spurious patterns. Given documents and their initial labels, we task humans with revising each document so that it (i) accords with a counterfactual target label; (ii) retains internal coherence; and (iii) avoids unnecessary changes. Interestingly, on sentiment analysis and natural language inference tasks, classifiers trained on original data fail on their counterfactually-revised counterparts and vice versa. Classifiers trained on combined datasets perform remarkably well, just shy of those specialized to either domain. While classifiers trained on either original or manipulated data alone are sensitive to spurious features (e.g., mentions of genre), models trained on the combined data are less sensitive to this signal. Both datasets are publicly available.

The Early Phase of Neural Network Training

- Jonathan Frankle, David J. Schwab, Ari S. Morcos
- abstract@open-review(Poster): Recent studies have shown that many important aspects of neural network learning take place within the very earliest iterations or epochs of training. For example, sparse, trainable sub-networks emerge (Frankle et al., 2019), gradient descent moves into a small subspace (Gur-Ari et al., 2018), and the network undergoes a critical period (Achille et al., 2019). Here we examine the changes that deep neural networks undergo during this early phase of training. We perform extensive measurements of the network state and its updates during these early iterations of training, and leverage the framework of Frankle et al. (2019) to quantitatively probe the weight distribution and its reliance on various aspects of the dataset. We find that, within this framework, deep networks are not robust to reinitializing with random weights while maintaining signs, and that weight distributions are highly non-independent even after only a few hundred iterations. Despite this, pre-training with blurred inputs or an auxiliary self-supervised task can approximate the changes in supervised networks, suggesting that these changes are label-agnostic, though labels significantly accelerate this process. Together, these results help to elucidate the network changes occurring during this pivotal initial period of learning.

RNNs Incrementally Evolving on an Equilibrium Manifold: A Panacea for Vanishing and Exploding Gradients?

- Anil Kag, Ziming Zhang, Venkatesh Saligrama
- abstract@open-review(Poster): Recurrent neural networks (RNNs) are particularly well-suited for modeling long-term dependencies in sequential data, but are notoriously hard to train because the error backpropagated in time either vanishes or explodes at an exponential rate. While a number of works attempt to mitigate this effect through gated recurrent units, skip-connections, parametric constraints and design choices, we propose a novel incremental RNN (iRNN), where hidden state vectors keep track of incremental changes, and as such approximate state-vector increments of Rosenblatt's (1962) continuous-time RNNs. iRNN exhibits identity gradients and is able to account for long-term dependencies (LTD). We show that our method is computationally efficient overcoming overheads of many existing methods that attempt to improve RNN training, while suffering no performance degradation. We demonstrate the utility of our approach with extensive experiments and show competitive performance against standard LSTMs on LTD and other non-LTD tasks.

Extreme Tensoring for Low-Memory Preconditioning

- Xinyi Chen, Naman Agarwal, Elad Hazan, Cyril Zhang, Yi Zhang
- abstract@open-review(Poster): State-of-the-art models are now trained with billions of parameters, reaching hardware limits in terms of memory consumption. This has created a recent demand for memory-efficient optimizers. To this end, we investigate the limits and performance tradeoffs of memory-efficient adaptively preconditioned gradient methods. We propose \emph{extreme tensoring} for high-dimensional stochastic optimization, showing that an optimizer needs very little memory to benefit from adaptive preconditioning. Our technique applies to arbitrary models (not necessarily with tensor-shaped parameters), and is accompanied by regret and convergence guarantees, which shed light on the tradeoffs between preconditioner quality and expressivity. On a large-scale NLP model, we reduce the optimizer memory overhead by three orders of magnitude, without degrading performance.

Non-Autoregressive Dialog State Tracking

- Hung Le, Richard Socher, Steven C.H. Hoi
- abstract@open-review(Poster): Recent efforts in Dialogue State Tracking (DST) for task-oriented dialogues have progressed toward open-vocabulary or generation-based approaches where the models can generate slot value candidates from the dialogue history itself. These approaches have shown good performance gain, especially in complicated dialogue domains with dynamic slot values. However, they fall short in two aspects: (1) they do not allow models to explicitly learn signals across domains and slots to detect potential dependencies among \textit{(domain, slot)} pairs; and (2) existing models follow auto-regressive approaches which incur high time cost when the dialogue evolves over multiple domains and multiple turns. In this paper, we propose a novel framework of Non-Autoregressive Dialog State Tracking (NADST) which can factor in potential dependencies among domains and slots to optimize the models towards better prediction of dialogue states as a complete set rather than separate slots. In particular, the non-autoregressive nature of our method not only enables decoding in parallel to significantly reduce the latency of DST for real-time dialogue response generation, but also detect dependencies among slots at token level in addition to slot and domain level. Our empirical results show that our model achieves the state-of-the-art joint accuracy across all domains on the MultiWOZ 2.1 corpus, and the latency of our model is an order of magnitude lower than the previous state of the art as the dialogue history extends over time.

Bayesian Meta Sampling for Fast Uncertainty Adaptation

- Zhenyi Wang, Yang Zhao, Ping Yu, Ruiyi Zhang, Changyou Chen

- abstract@open-review(Poster): Meta learning has been making impressive progress for fast model adaptation. However, limited work has been done on learning fast uncertainty adaption for Bayesian modeling. In this paper, we propose to achieve the goal by placing meta learning on the space of probability measures, inducing the concept of meta sampling for fast uncertainty adaption. Specifically, we propose a Bayesian meta sampling framework consisting of two main components: a meta sampler and a sample adapter. The meta sampler is constructed by adopting a neural-inverse-autoregressive-flow (NIAF) structure, a variant of the recently proposed neural autoregressive flows, to efficiently generate meta samples to be adapted. The sample adapter moves meta samples to task-specific samples, based on a newly proposed and general Bayesian sampling technique, called optimal-transport Bayesian sampling. The combination of the two components allows a simple learning procedure for the meta sampler to be developed, which can be efficiently optimized via standard back-propagation. Extensive experimental results demonstrate the efficiency and effectiveness of the proposed framework, obtaining better sample quality and faster uncertainty adaption compared to related methods.

Harnessing Structures for Value-Based Planning and Reinforcement Learning

- Yuzhe Yang, Guo Zhang, Zhi Xu, Dina Katabi
- abstract@open-review(Talk): Value-based methods constitute a fundamental methodology in planning and deep reinforcement learning (RL). In this paper, we propose to exploit the underlying structures of the state-action value function, i.e., Q function, for both planning and deep RL. In particular, if the underlying system dynamics lead to some global structures of the Q function, one should be capable of inferring the function better by leveraging such structures. Specifically, we investigate the low-rank structure, which widely exists for big data matrices. We verify empirically the existence of low-rank Q functions in the context of control and deep RL tasks. As our key contribution, by leveraging Matrix Estimation (ME) techniques, we propose a general framework to exploit the underlying low-rank structure in Q functions. This leads to a more efficient planning procedure for classical control, and additionally, a simple scheme that can be applied to value-based RL techniques to consistently achieve better performance on "low-rank" tasks. Extensive experiments on control tasks and Atari games confirm the efficacy of our approach.

Economy Statistical Recurrent Units For Inferring Nonlinear Granger Causality

- Saurabh Khanna, Vincent Y. F. Tan
- abstract@open-review(Poster): Granger causality is a widely-used criterion for analyzing interactions in large-scale networks. As most physical interactions are inherently nonlinear, we consider the problem of inferring the existence of pairwise Granger causality between nonlinearly interacting stochastic processes from their time series measurements. Our proposed approach relies on modeling the embedded nonlinearities in the measurements using a component-wise time series prediction model based on Statistical Recurrent Units (SRUs). We make a case that the network topology of Granger causal relations is directly inferrable from a structured sparse estimate of the internal parameters of the SRU networks trained to predict the processes' time series measurements. We propose a variant of SRU, called economy-SRU, which, by design has considerably fewer trainable parameters, and therefore less prone to overfitting. The economy-SRU computes a low-dimensional sketch of its high-dimensional hidden state in the form of random projections to generate the feedback for its recurrent processing. Additionally, the internal weight parameters of the economy-SRU are strategically regularized in a group-wise manner to facilitate the proposed network in extracting meaningful predictive features that are highly time-localized to mimic real-world causal events. Extensive experiments are carried out to demonstrate that the proposed economy-SRU based time series prediction model outperforms the MLP, LSTM and attention-gated CNN-based time series models considered previously for inferring Granger causality.

MEMO: A Deep Network for Flexible Combination of Episodic Memories

- Andrea Banino, Adrià Puigdomènech Badia, Raphael Köster, Martin J. Chadwick, Vinicius Zambaldi, Demis Hassabis, Caswell Barry, Matthew Botvinick, Dharshan Kumaran, Charles Blundell
- abstract@open-review(Poster): Recent research developing neural network architectures with external memory have often used the benchmark bAbI question and answering dataset which provides a challenging number of tasks requiring reasoning. Here we employed a classic associative inference task from the human neuroscience literature in order to more carefully probe the reasoning capacity of existing memory-augmented architectures. This task is thought to capture the essence of reasoning -- the appreciation of distant relationships among elements distributed across multiple facts or memories. Surprisingly, we found that current architectures struggle to reason over long distance associations. Similar results were obtained on a more complex task involving finding the shortest path between nodes in a path. We therefore developed a novel architecture, MEMO, endowed with the capacity to reason over longer distances. This was accomplished with the addition of two novel components. First, it introduces a separation between memories/facts stored in external memory and the items that comprise these facts in external memory. Second, it makes use of an adaptive retrieval mechanism, allowing a variable number of 'memory hops' before the answer is produced. MEMO is capable of solving our novel reasoning tasks, as well as all 20 tasks in bAbI.

Probabilistic Connection Importance Inference and Lossless Compression of Deep Neural Networks

- Xin Xing, Long Sha, Pengyu Hong, Zuofeng Shang, Jun S. Liu
- abstract@open-review(Poster): Deep neural networks (DNNs) can be huge in size, requiring a considerable a mount of energy and computational resources to operate, which limits their applications in numerous scenarios. It is thus of interest to compress DNNs while maintaining their performance levels. We here propose a probabilistic importance inference approach for pruning DNNs. Specifically, we test the significance of the relevance of a connection in a DNN to the DNN's outputs using a nonparemtric scoring testand keep only those significant ones. Experimental results show that the proposed approach achieves better lossless compression rates than existing techniques

Coherent Gradients: An Approach to Understanding Generalization in Gradient Descent-based Optimization

- Satrajit Chatterjee
- abstract@open-review(Poster): An open question in the Deep Learning community is why neural networks trained with Gradient Descent generalize well on real datasets even though they are capable of fitting random data. We propose an approach to answering this question based on a hypothesis about the dynamics of gradient descent that we call Coherent Gradients: Gradients from similar examples are similar and so the overall gradient is stronger in certain directions where these reinforce each other. Thus changes to the network parameters during training are biased towards those that (locally) simultaneously benefit many examples when such similarity exists. We support this hypothesis with heuristic arguments and perturbative experiments and outline how this can explain several common empirical observations about Deep Learning. Furthermore, our analysis is not just descriptive, but prescriptive. It suggests a natural modification to gradient descent that can greatly reduce overfitting.

Jelly Bean World: A Testbed for Never-Ending Learning

- Emmanouil Antonios Platanios, Abulhair Saparov, Tom Mitchell
- abstract@open-review(Poster): Machine learning has shown growing success in recent years. However, current machine learning systems are highly specialized, trained for particular problems or domains, and typically on a single narrow dataset. Human learning, on the other hand, is highly general and adaptable. Never-ending learning is a machine learning paradigm that aims to bridge this gap, with the goal of encouraging researchers to design machine learning systems that can learn to perform a wider variety of inter-related tasks in more complex environments. To date, there is no environment or testbed to facilitate the development and evaluation of never-ending learning systems. To this end, we propose the Jelly Bean World testbed. The Jelly Bean World allows experimentation over two-dimensional grid worlds which are filled with items and in which agents can navigate. This testbed provides environments that are sufficiently complex and where more generally intelligent algorithms ought to perform better than current state-of-the-art reinforcement learning approaches. It does so by producing non-stationary environments and facilitating experimentation with multi-task, multi-agent,

multi-modal, and curriculum learning settings. We hope that this new freely-available software will prompt new research and interest in the development and evaluation of never-ending learning systems and more broadly, general intelligence systems.

Learning from Explanations with Neural Execution Tree

- Ziqi Wang, *Yujia Qin*, Wenxuan Zhou, Jun Yan, Qinyuan Ye, Leonardo Neves, Zhiyuan Liu, Xiang Ren
- abstract@open-review(Poster): While deep neural networks have achieved impressive performance on a range of NLP tasks, these data-hungry models heavily rely on labeled data, which restricts their applications in scenarios where data annotation is expensive. Natural language (NL) explanations have been demonstrated very useful additional supervision, which can provide sufficient domain knowledge for generating more labeled data over new instances, while the annotation time only doubles. However, directly applying them for augmenting model learning encounters two challenges: (1) NL explanations are unstructured and inherently compositional, which asks for a modularized model to represent their semantics, (2) NL explanations often have large numbers of linguistic variants, resulting in low recall and limited generalization ability. In this paper, we propose a novel Neural Execution Tree (NExT) framework to augment training data for text classification using NL explanations. After transforming NL explanations into executable logical forms by semantic parsing, NExT generalizes different types of actions specified by the logical forms for labeling data instances, which substantially increases the coverage of each NL explanation. Experiments on two NLP tasks (relation extraction and sentiment analysis) demonstrate its superiority over baseline methods. Its extension to multi-hop question answering achieves performance gain with light annotation effort.

Discovering Motor Programs by Recomposing Demonstrations

- Tanmay Shankar, Shubham Tulsiani, Lerrel Pinto, Abhinav Gupta
- abstract@open-review(Poster): In this paper, we present an approach to learn recomposable motor primitives across large-scale and diverse manipulation demonstrations. Current approaches to decomposing demonstrations into primitives often assume manually defined primitives and bypass the difficulty of discovering these primitives. On the other hand, approaches in primitive discovery put restrictive assumptions on the complexity of a primitive, which limit applicability to narrow tasks. Our approach attempts to circumvent these challenges by jointly learning both the underlying motor primitives and recomposing these primitives to form the original demonstration. Through constraints on both the parsimony of primitive decomposition and the simplicity of a given primitive, we are able to learn a diverse set of motor primitives, as well as a coherent latent representation for these primitives. We demonstrate both qualitatively and quantitatively, that our learned primitives capture semantically meaningful aspects of a demonstration. This allows us to compose these primitives in a hierarchical reinforcement learning setup to efficiently solve robotic manipulation tasks like reaching and pushing. Our results may be viewed at https://sites.google.com/view/discovering-motor-programs.

Convergence of Gradient Methods on Bilinear Zero-Sum Games

- Guojun Zhang, Yaoliang Yu
- abstract@open-review(Poster): Min-max formulations have attracted great attention in the ML community due to the rise of deep generative models and adversarial methods, while understanding the dynamics of gradient algorithms for solving such formulations has remained a grand challenge. As a first step, we restrict to bilinear zero-sum games and give a systematic analysis of popular gradient updates, for both simultaneous and alternating versions. We provide exact conditions for their convergence and find the optimal parameter setup and convergence rates. In particular, our results offer formal evidence that alternating updates converge "better" than simultaneous ones.

Composing Task-Agnostic Policies with Deep Reinforcement Learning

- Ahmed H. Qureshi, Jacob J. Johnson, Yuzhe Qin, Taylor Henderson, Byron Boots, Michael C. Yip
- abstract@open-review(Poster): The composition of elementary behaviors to solve challenging transfer learning problems is one of the key elements in building intelligent machines. To date, there has been plenty of work on learning task-specific policies or skills but almost no focus on composing necessary, task-agnostic skills to find a solution to new problems. In this paper, we propose a novel deep reinforcement learning-based skill transfer and composition method that takes the agent's primitive policies to solve unseen tasks. We evaluate our method in difficult cases where training policy through standard reinforcement learning (RL) or even hierarchical RL is either not feasible or exhibits high sample complexity. We show that our method not only transfers skills to new problem settings but also solves the challenging environments requiring both task planning and motion control with high data efficiency.

The Local Elasticity of Neural Networks

- Hangfeng He, Weijie Su
- abstract@open-review(Poster): This paper presents a phenomenon in neural networks that we refer to as local elasticity. Roughly speaking, a classifier is said to be locally elastic if its prediction at a feature vector x' is not significantly perturbed, after the classifier is updated via stochastic gradient descent at a (labeled) feature vector x that is dissimilar to x' in a certain sense. This phenomenon is shown to persist for neural networks with nonlinear activation functions through extensive simulations on real-life and synthetic datasets, whereas this is not observed in linear classifiers. In addition, we offer a geometric interpretation of local elasticity using the neural tangent kernel (Jacot et al., 2018). Building on top of local elasticity, we obtain pairwise similarity measures between feature vectors, which can be used for clustering in conjunction with K-means. The effectiveness of the clustering algorithm on the MNIST and CIFAR-10 datasets in turn corroborates the hypothesis of local elasticity of neural networks on real-life data. Finally, we discuss some implications of local elasticity to shed light on several intriguing aspects of deep neural networks.

Gradient-Based Neural DAG Learning

- Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, Simon Lacoste-Julien
- abstract@open-review(Poster): We propose a novel score-based approach to learning a directed acyclic graph (DAG) from observational data. We adapt a recently proposed continuous constrained optimization formulation to allow for nonlinear relationships between variables using neural networks. This extension allows to model complex interactions while avoiding the combinatorial nature of the problem. In addition to comparing our method to existing continuous optimization methods, we provide missing empirical comparisons to nonlinear greedy search methods. On both synthetic and real-world data sets, this new method outperforms current continuous methods on most tasks while being competitive with existing greedy search methods on important metrics for causal inference.

Composition-based Multi-Relational Graph Convolutional Networks

- Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, Partha Talukdar
- abstract@open-review(Poster): Graph Convolutional Networks (GCNs) have recently been shown to be quite successful in modeling graph-structured data. However, the primary focus has been on handling simple undirected graphs. Multi-relational graphs are a more general and prevalent form of graphs where each edge has a label and direction associated with it. Most of the existing approaches to handle such graphs suffer from over-parameterization and are restricted to learning representations of nodes only. In this paper, we propose CompGCN, a novel Graph Convolutional framework which jointly embeds both nodes and relations in a relational graph. CompGCN leverages a variety of entity-relation composition operations from Knowledge Graph Embedding techniques and scales with the number of relations. It also generalizes several of the existing multi-relational GCN methods. We evaluate our

proposed method on multiple tasks such as node classification, link prediction, and graph classification, and achieve demonstrably superior results. We make the source code of CompGCN available to foster reproducible research.

Capsules with Inverted Dot-Product Attention Routing

- Yao-Hung Hubert Tsai, Nitish Srivastava, Hanlin Goh, Ruslan Salakhutdinov
- abstract@open-review(Poster): We introduce a new routing algorithm for capsule networks, in which a child capsule is routed to a parent based only on agreement between the parent's state and the child's vote. The new mechanism 1) designs routing via inverted dot-product attention; 2) imposes Layer Normalization as normalization; and 3) replaces sequential iterative routing with concurrent iterative routing. When compared to previously proposed routing algorithms, our method improves performance on benchmark datasets such as CIFAR-10 and CIFAR-100, and it performs at-par with a powerful CNN (ResNet-18) with 4x fewer parameters. On a different task of recognizing digits from overlayed digit images, the proposed capsule model performs favorably against CNNs given the same number of layers and neurons per layer. We believe that our work raises the possibility of applying capsule networks to complex real-world tasks.

FSNet: Compression of Deep Convolutional Neural Networks by Filter Summary

- Yingzhen Yang, Jiahui Yu, Nebojsa Jojic, Jun Huan, Thomas S. Huang
- abstract@open-review(Poster): We present a novel method of compression of deep Convolutional Neural Networks (CNNs) by weight sharing through a new representation of convolutional filters. The proposed method reduces the number of parameters of each convolutional layer by learning a 1D vector termed Filter Summary (FS). The convolutional filters are located in FS as overlapping 1D segments, and nearby filters in FS share weights in their overlapping regions in a natural way. The resultant neural network based on such weight sharing scheme, termed Filter Summary CNNs or FSNet, has a FS in each convolution layer instead of a set of independent filters in the conventional convolution layer. FSNet has the same architecture as that of the baseline CNN to be compressed, and each convolution layer of FSNet has the same number of filters from FS as that of the basline CNN in the forward process. With compelling computational acceleration ratio, the parameter space of FSNet is much smaller than that of the baseline CNN. In addition, FSNet is quantization friendly. FSNet with weight quantization leads to even higher compression ratio without noticeable performance loss. We further propose Differentiable FSNet where the way filters share weights is learned in a differentiable and end-to-end manner. Experiments demonstrate the effectiveness of FSNet in compression of CNNs for computer vision tasks including image classification and object detection, and the effectiveness of DFSNet is evidenced by the task of Neural Architecture Search.

On the Need for Topology-Aware Generative Models for Manifold-Based Defenses

- Uyeong Jang, Susmit Jha, Somesh Jha
- abstract@open-review(Poster): ML algorithms or models, especially deep neural networks (DNNs), have shown significant promise in several areas. However, recently researchers have demonstrated that ML algorithms, especially DNNs, are vulnerable to adversarial examples (slightly perturbed samples that cause mis-classification). Existence of adversarial examples has hindered deployment of ML algorithms in safety-critical sectors, such as security. Several defenses for adversarial examples exist in the literature. One of the important classes of defenses are manifold-based defenses, where a sample is "pulled back" into the data manifold before classifying. These defenses rely on the manifold assumption (data lie in a manifold of lower dimension than the input space). These defenses use a generative model to approximate the input distribution. This paper asks the following question: do the generative models used in manifold-based defenses need to be topology-aware? Our paper suggests the answer is yes. We provide theoretical and empirical evidence to support our claim.

Neural Execution of Graph Algorithms

- Petar Veličković, Rex Ying, Matilde Padovano, Raia Hadsell, Charles Blundell
- abstract@open-review(Poster): Graph Neural Networks (GNNs) are a powerful representational tool for solving problems on graph-structured inputs. In almost all cases so far, however, they have been applied to directly recovering a final solution from raw inputs, without explicit guidance on how to structure their problem-solving. Here, instead, we focus on learning in the space of algorithms: we train several state-of-the-art GNN architectures to imitate individual steps of classical graph algorithms, parallel (breadth-first search, Bellman-Ford) as well as sequential (Prim's algorithm). As graph algorithms usually rely on making discrete decisions within neighbourhoods, we hypothesise that maximisation-based message passing neural networks are best-suited for such objectives, and validate this claim empirically. We also demonstrate how learning in the space of algorithms can yield new opportunities for positive transfer between tasks---showing how learning a shortest-path algorithm can be substantially improved when simultaneously learning a reachability algorithm.

BERTScore: Evaluating Text Generation with BERT

- Tianyi Zhang*, *Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, Yoav Artzi
- abstract@open-review(Poster): We propose BERTScore, an automatic evaluation metric for text generation. Analogously to common metrics, BERTScore computes a similarity score for each token in the candidate sentence with each token in the reference sentence. However, instead of exact matches, we compute token similarity using contextual embeddings. We evaluate using the outputs of 363 machine translation and image captioning systems. BERTScore correlates better with human judgments and provides stronger model selection performance than existing metrics. Finally, we use an adversarial paraphrase detection task and show that BERTScore is more robust to challenging examples compared to existing metrics.

Augmenting Non-Collaborative Dialog Systems with Explicit Semantic and Strategic Dialog History

- Yiheng Zhou, Yulia Tsvetkov, Alan W Black, Zhou Yu
- abstract@open-review(Poster): We study non-collaborative dialogs, where two agents have a conflict of interest but must strategically communicate to reach an agreement (e.g., negotiation). This setting poses new challenges for modeling dialog history because the dialog's outcome relies not only on the semantic intent, but also on tactics that convey the intent. We propose to model both semantic and tactic history using finite state transducers (FSTs). Unlike RNN, FSTs can explicitly represent dialog history through all the states traversed, facilitating interpretability of dialog structure. We train FSTs on a set of strategies and tactics used in negotiation dialogs. The trained FSTs show plausible tactic structure and can be generalized to other non-collaborative domains (e.g., persuasion). We evaluate the FSTs by incorporating them in an automated negotiating system that attempts to sell products and a persuasion system that persuades people to donate to a charity. Experiments show that explicitly modeling both semantic and tactic history is an effective way to improve both dialog policy planning and generation performance.

Prediction, Consistency, Curvature: Representation Learning for Locally-Linear Control

- Nir Levine, Yinlam Chow, Rui Shu, Ang Li, Mohammad Ghavamzadeh, Hung Bui
- abstract@open-review(Poster): Many real-world sequential decision-making problems can be formulated as optimal control with high-dimensional observations and unknown dynamics. A promising approach is to embed the high-dimensional observations into a lower-dimensional latent representation space, estimate the latent dynamics model, then utilize this model for control in the latent space. An important open question is how to learn a representation that is amenable to existing control algorithms? In this paper, we focus on learning representations for locally-linear control algorithms,

such as iterative LQR (iLQR). By formulating and analyzing the representation learning problem from an optimal control perspective, we establish three underlying principles that the learned representation should comprise: 1) accurate prediction in the observation space, 2) consistency between latent and observation space dynamics, and 3) low curvature in the latent space transitions. These principles naturally correspond to a loss function that consists of three terms: prediction, consistency, and curvature (PCC). Crucially, to make PCC tractable, we derive an amortized variational bound for the PCC loss function. Extensive experiments on benchmark domains demonstrate that the new variational-PCC learning algorithm benefits from significantly more stable and reproducible training, and leads to superior control performance. Further ablation studies give support to the importance of all three PCC components for learning a good latent space for control.

GraphZoom: A Multi-level Spectral Approach for Accurate and Scalable Graph Embedding

- Chenhui Deng, Zhiqiang Zhao, Yongyu Wang, Zhiru Zhang, Zhuo Feng
- abstract@open-review(Talk): Graph embedding techniques have been increasingly deployed in a multitude of different applications that involve learning on non-Euclidean data. However, existing graph embedding models either fail to incorporate node attribute information during training or suffer from node attribute noise, which compromises the accuracy. Moreover, very few of them scale to large graphs due to their high computational complexity and memory usage. In this paper we propose GraphZoom, a multi-level framework for improving both accuracy and scalability of unsupervised graph embedding algorithms. GraphZoom first performs graph fusion to generate a new graph that effectively encodes the topology of the original graph and the node attribute information. This fused graph is then repeatedly coarsened into much smaller graphs by merging nodes with high spectral similarities. GraphZoom allows any existing embedding methods to be applied to the coarsened graph, before it progressively refine the embeddings obtained at the coarsest level to increasingly finer graphs. We have evaluated our approach on a number of popular graph datasets for both transductive and inductive tasks. Our experiments show that GraphZoom can substantially increase the classification accuracy and significantly accelerate the entire graph embedding process by up to $40.8 \times$, when compared to the state-of-the-art unsupervised embedding methods.

Graph Constrained Reinforcement Learning for Natural Language Action Spaces

- Prithviraj Ammanabrolu, Matthew Hausknecht
- abstract@open-review(Poster): Interactive Fiction games are text-based simulations in which an agent interacts with the world purely through natural language. They are ideal environments for studying how to extend reinforcement learning agents to meet the challenges of natural language understanding, partial observability, and action generation in combinatorially-large text-based action spaces. We present KG-A2C, an agent that builds a dynamic knowledge graph while exploring and generates actions using a template-based action space. We contend that the dual uses of the knowledge graph to reason about game state and to constrain natural language generation are the keys to scalable exploration of combinatorially large natural language actions. Results across a wide variety of IF games show that KG-A2C outperforms current IF agents despite the exponential increase in action space size.

Towards Fast Adaptation of Neural Architectures with Meta Learning

- Dongze Lian, Yin Zheng, Yintao Xu, Yanxiong Lu, Leyu Lin, Peilin Zhao, Junzhou Huang, Shenghua Gao
- abstract@open-review(Poster): Recently, Neural Architecture Search (NAS) has been successfully applied to multiple artificial intelligence areas and shows better performance compared with hand-designed networks. However, the existing NAS methods only target a specific task. Most of them usually do well in searching an architecture for single task but are troublesome for multiple datasets or multiple tasks. Generally, the architecture for a new task is either searched from scratch, which is neither efficient nor flexible enough for practical application scenarios, or borrowed from the ones searched on other tasks, which might be not optimal. In order to tackle the transferability of NAS and conduct fast adaptation of neural architectures, we propose a novel Transferable Neural Architecture Search method based on meta-learning in this paper, which is termed as T-NAS. T-NAS learns a meta-architecture that is able to adapt to a new task quickly through a few gradient steps, which makes the transferred architecture suitable for the specific task. Extensive experiments show that T-NAS achieves state-of-the-art performance in few-shot learning and comparable performance in supervised learning but with 50x less searching cost, which demonstrates the effectiveness of our method.

Variational Hetero-Encoder Randomized GANs for Joint Image-Text Modeling

- Hao Zhang, Bo Chen, Long Tian, Zhengjue Wang, Mingyuan Zhou
- abstract@open-review(Poster): For bidirectional joint image-text modeling, we develop variational hetero-encoder (VHE) randomized generative adversarial network (GAN), a versatile deep generative model that integrates a probabilistic text decoder, probabilistic image encoder, and GAN into a coherent end-to-end multi-modality learning framework. VHE randomized GAN (VHE-GAN) encodes an image to decode its associated text, and feeds the variational posterior as the source of randomness into the GAN image generator. We plug three off-the-shelf modules, including a deep topic model, a ladder-structured image encoder, and StackGAN++, into VHE-GAN, which already achieves competitive performance. This further motivates the development of VHE-raster-scan-GAN that generates photo-realistic images in not only a multi-scale low-to-high-resolution manner, but also a hierarchical-semantic coarse-to-fine fashion. By capturing and relating hierarchical semantic and visual concepts with end-to-end training, VHE-raster-scan-GAN achieves state-of-the-art performance in a wide variety of image-text multi-modality learning and generation tasks.

Asymptotics of Wide Networks from Feynman Diagrams

- Ethan Dyer, Guy Gur-Ari
- abstract@open-review(Spotlight): Understanding the asymptotic behavior of wide networks is of considerable interest. In this work, we present a general method for analyzing this large width behavior. The method is an adaptation of Feynman diagrams, a standard tool for computing multivariate Gaussian integrals. We apply our method to study training dynamics, improving existing bounds and deriving new results on wide network evolution during stochastic gradient descent. Going beyond the strict large width limit, we present closed-form expressions for higher-order terms governing wide network training, and test these predictions empirically.

Symplectic Recurrent Neural Networks

- Zhengdao Chen, Jianyu Zhang, Martin Arjovsky, Léon Bottou
- abstract@open-review(Spotlight): We propose Symplectic Recurrent Neural Networks (SRNNs) as learning algorithms that capture the dynamics of physical systems from observed trajectories. SRNNs model the Hamiltonian function of the system by a neural networks, and leverage symplectic integration, multiple-step training and initial state optimization to address the challenging numerical issues associated with Hamiltonian systems. We show SRNNs succeed reliably on complex and noisy Hamiltonian systems. Finally, we show how to augment the SRNN integration scheme in order to handle stiff dynamical systems such as bouncing billiards.

Higher-Order Function Networks for Learning Composable 3D Object Representations

- Eric Mitchell, Selim Engin, Volkan Isler, Daniel D Lee
- abstract@open-review(Poster): We present a new approach to 3D object representation where a neural network encodes the geometry of an object directly into the weights and biases of a second 'mapping' network. This mapping network can be used to reconstruct an object by applying its encoded

transformation to points randomly sampled from a simple geometric space, such as the unit sphere. We study the effectiveness of our method through various experiments on subsets of the ShapeNet dataset. We find that the proposed approach can reconstruct encoded objects with accuracy equal to or exceeding state-of-the-art methods with orders of magnitude fewer parameters. Our smallest mapping network has only about 7000 parameters and shows reconstruction quality on par with state-of-the-art object decoder architectures with millions of parameters. Further experiments on feature mixing through the composition of learned functions show that the encoding captures a meaningful subspace of objects.

Neural Module Networks for Reasoning over Text

- Nitish Gupta, Kevin Lin, Dan Roth, Sameer Singh, Matt Gardner
- abstract@open-review(Poster): Answering compositional questions that require multiple steps of reasoning against text is challenging, especially when they involve discrete, symbolic operations. Neural module networks (NMNs) learn to parse such questions as executable programs composed of learnable modules, performing well on synthetic visual QA domains. However, we find that it is challenging to learn these models for non-synthetic questions on open-domain text, where a model needs to deal with the diversity of natural language and perform a broader range of reasoning. We extend NMNs by: (a) introducing modules that reason over a paragraph of text, performing symbolic reasoning (such as arithmetic, sorting, counting) over numbers and dates in a probabilistic and differentiable manner; and (b) proposing an unsupervised auxiliary loss to help extract arguments associated with the events in text. Additionally, we show that a limited amount of heuristically-obtained question program and intermediate module output supervision provides sufficient inductive bias for accurate learning. Our proposed model significantly outperforms state-of-the-art models on a subset of the DROP dataset that poses a variety of reasoning challenges that are covered by our modules.

Learning to Plan in High Dimensions via Neural Exploration-Exploitation Trees

- Binghong Chen, Bo Dai, Qinjie Lin, Guo Ye, Han Liu, Le Song
- abstract@open-review(Spotlight): We propose a meta path planning algorithm named \emph{Neural Exploration-Exploitation Trees~(NEXT)} for learning from prior experience for solving new path planning problems in high dimensional continuous state and action spaces. Compared to more classical sampling-based methods like RRT, our approach achieves much better sample efficiency in high-dimensions and can benefit from prior experience of planning in similar environments. More specifically, NEXT exploits a novel neural architecture which can learn promising search directions from problem structures. The learned prior is then integrated into a UCB-type algorithm to achieve an online balance between \emph{exploration} and \emph{exploitation} when solving a new problem. We conduct thorough experiments to show that NEXT accomplishes new planning problems with more compact search trees and significantly outperforms state-of-the-art methods on several benchmarks.

Improved memory in recurrent neural networks with sequential non-normal dynamics

- Emin Orhan, Xaq Pitkow
- abstract@open-review(Poster): Training recurrent neural networks (RNNs) is a hard problem due to degeneracies in the optimization landscape, a problem also known as vanishing/exploding gradients. Short of designing new RNN architectures, previous methods for dealing with this problem usually boil down to orthogonalization of the recurrent dynamics, either at initialization or during the entire training period. The basic motivation behind these methods is that orthogonal transformations are isometries of the Euclidean space, hence they preserve (Euclidean) norms and effectively deal with vanishing/exploding gradients. However, this ignores the crucial effects of non-linearity and noise. In the presence of a non-linearity, orthogonal transformations no longer preserve norms, suggesting that alternative transformations might be better suited to non-linear networks. Moreover, in the presence of noise, norm preservation itself ceases to be the ideal objective. A more sensible objective is maximizing the signal-to-noise ratio (SNR) of the propagated signal instead. Previous work has shown that in the linear case, recurrent networks that maximize the SNR display strongly non-normal, sequential dynamics and orthogonal networks are highly suboptimal by this measure. Motivated by this finding, here we investigate the potential of non-normal RNNs, i.e. RNNs with a non-normal recurrent connectivity matrix, in sequential processing tasks. Our experimental results show that non-normal RNNs outperform their orthogonal counterparts in a diverse range of benchmarks. We also find evidence for increased non-normality and hidden chain-like feedforward motifs in trained RNNs initialized with orthogonal recurrent connectivity matrices.

Learn to Explain Efficiently via Neural Logic Inductive Learning

- Yuan Yang, Le Song
- abstract@open-review(Poster): The capability of making interpretable and self-explanatory decisions is essential for developing responsible machine learning systems. In this work, we study the learning to explain the problem in the scope of inductive logic programming (ILP). We propose Neural Logic Inductive Learning (NLIL), an efficient differentiable ILP framework that learns first-order logic rules that can explain the patterns in the data. In experiments, compared with the state-of-the-art models, we find NLIL is able to search for rules that are x10 times longer while remaining x3 times faster. We also show that NLIL can scale to large image datasets, i.e. Visual Genome, with 1M entities.

Improving Neural Language Generation with Spectrum Control

- Lingxiao Wang, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang, Quanquan Gu
- abstract@open-review(Poster): Recent Transformer-based models such as Transformer-XL and BERT have achieved huge success on various natural language processing tasks. However, contextualized embeddings at the output layer of these powerful models tend to degenerate and occupy an anisotropic cone in the vector space, which is called the representation degeneration problem. In this paper, we propose a novel spectrum control approach to address this degeneration problem. The core idea of our method is to directly guide the spectra training of the output embedding matrix with a slow-decaying singular value prior distribution through a reparameterization framework. We show that our proposed method encourages isotropy of the learned word representations while maintains the modeling power of these contextual neural models. We further provide a theoretical analysis and insight on the benefit of modeling singular value distribution. We demonstrate that our spectrum control method outperforms the state-of-the-art Transformer-XL modeling for language model, and various Transformer-based models for machine translation, on common benchmark datasets for these tasks.

Span Recovery for Deep Neural Networks with Applications to Input Obfuscation

- Rajesh Jayaram, David P. Woodruff, Qiuyi Zhang
- abstract@open-review(Poster): The tremendous success of deep neural networks has motivated the need to better understand the fundamental properties of these networks, but many of the theoretical results proposed have only been for shallow networks. In this paper, we study an important primitive for understanding the meaningful input space of a deep network: span recovery. For $k<n$, let $\mathbf{A} \in \mathbb{R}^{k \times n}$ be the innermost weight matrix of an arbitrary feed forward neural network $M: \mathbb{R}^n \to \mathbb{R}$, so $M(x)$ can be written as $M(x) = \sigma(\mathbf{A} x)$, for some network $\sigma: \mathbb{R}^k \to \mathbb{R}$. The goal is then to recover the row span of $\mathbf{A}$ given only oracle access to the value of $M(x)$. We show that if M is a multi-layered network with ReLU activation functions, then partial recovery is possible: namely, we can provably recover $k/2$ linearly independent vectors in the row span of $\mathbf{A}$ using poly(n) non-adaptive queries to $M(x)$. Furthermore, if M has differentiable activation functions, we demonstrate that \textit{full} span recovery is possible even when the output is first passed through a sign or $0/1$ thresholding function; in this case our algorithm is adaptive. Empirically, we confirm that full span recovery is not always possible, but only for unrealistically thin layers. For reasonably wide networks, we obtain full span recovery on both random networks and networks trained on MNIST data. Furthermore, we demonstrate the utility of span recovery as an attack by inducing neural networks to misclassify data obfuscated by controlled random noise as sensical inputs.

Oblique Decision Trees from Derivatives of ReLU Networks

- Guang-He Lee, Tommi S. Jaakkola
- abstract@open-review(Poster): We show how neural models can be used to realize piece-wise constant functions such as decision trees. The proposed architecture, which we call locally constant networks, builds on ReLU networks that are piece-wise linear and hence their associated gradients with respect to the inputs are locally constant. We formally establish the equivalence between the classes of locally constant networks and decision trees. Moreover, we highlight several advantageous properties of locally constant networks, including how they realize decision trees with parameter sharing across branching / leaves. Indeed, only M neurons suffice to implicitly model an oblique decision tree with 2^M leaf nodes. The neural representation also enables us to adopt many tools developed for deep networks (e.g., DropConnect (Wan et al., 2013)) while implicitly training decision trees. We demonstrate that our method outperforms alternative techniques for training oblique decision trees in the context of molecular property classification and regression tasks.

Precision Gating: Improving Neural Network Efficiency with Dynamic Dual-Precision Activations

- Yichi Zhang, Ritchie Zhao, Weizhe Hua, Nayun Xu, G. Edward Suh, Zhiru Zhang
- abstract@open-review(Poster): We propose precision gating (PG), an end-to-end trainable dynamic dual-precision quantization technique for deep neural networks. PG computes most features in a low precision and only a small proportion of important features in a higher precision to preserve accuracy. The proposed approach is applicable to a variety of DNN architectures and significantly reduces the computational cost of DNN execution with almost no accuracy loss. Our experiments indicate that PG achieves excellent results on CNNs, including statically compressed mobile-friendly networks such as ShuffleNet. Compared to the state-of-the-art prediction-based quantization schemes, PG achieves the same or higher accuracy with 2.4× less compute on ImageNet. PG furthermore applies to RNNs. Compared to 8-bit uniform quantization, PG obtains a 1.2% improvement in perplexity per word with 2.7× computational cost reduction on LSTM on the Penn Tree Bank dataset.

PAC Confidence Sets for Deep Neural Networks via Calibrated Prediction

- Sangdon Park, Osbert Bastani, Nikolai Matni, Insup Lee
- abstract@open-review(Poster): We propose an algorithm combining calibrated prediction and generalization bounds from learning theory to construct confidence sets for deep neural networks with PAC guarantees---i.e., the confidence set for a given input contains the true label with high probability. We demonstrate how our approach can be used to construct PAC confidence sets on ResNet for ImageNet, a visual object tracking model, and a dynamics model for the half-cheetah reinforcement learning problem.

DD-PPO: Learning Near-Perfect PointGoal Navigators from 2.5 Billion Frames

- Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, Dhruv Batra
- abstract@open-review(Poster): We present Decentralized Distributed Proximal Policy Optimization (DD-PPO), a method for distributed reinforcement learning in resource-intensive simulated environments. DD-PPO is distributed (uses multiple machines), decentralized (lacks a centralized server), and synchronous (no computation is ever "stale"), making it conceptually simple and easy to implement. In our experiments on training virtual robots to navigate in Habitat-Sim, DD-PPO exhibits near-linear scaling -- achieving a speedup of 107x on 128 GPUs over a serial implementation. We leverage this scaling to train an agent for 2.5 Billion steps of experience (the equivalent of 80 years of human experience) -- over 6 months of GPU-time training in under 3 days of wall-clock time with 64 GPUs.

This massive-scale training not only sets the state of art on Habitat Autonomous Navigation Challenge 2019, but essentially "solves" the task -- near-perfect autonomous navigation in an unseen environment without access to a map, directly from an RGB-D camera and a GPS+Compass sensor. Fortuitously, error vs computation exhibits a power-law-like distribution; thus, 90% of peak performance is obtained relatively early (at 100 million steps) and relatively cheaply (under 1 day with 8 GPUs). Finally, we show that the scene understanding and navigation policies learned can be transferred to other navigation tasks -- the analog of "ImageNet pre-training + task-specific fine-tuning" for embodied AI. Our model outperforms ImageNet pre-trained CNNs on these transfer tasks and can serve as a universal resource (all models and code are publicly available).

Learning to Learn by Zeroth-Order Oracle

- Yangjun Ruan, Yuanhao Xiong, Sashank Reddi, Sanjiv Kumar, Cho-Jui Hsieh
- abstract@open-review(Poster): In the learning to learn (L2L) framework, we cast the design of optimization algorithms as a machine learning problem and use deep neural networks to learn the update rules. In this paper, we extend the L2L framework to zeroth-order (ZO) optimization setting, where no explicit gradient information is available. Our learned optimizer, modeled as recurrent neural network (RNN), first approximates gradient by ZO gradient estimator and then produces parameter update utilizing the knowledge of previous iterations. To reduce high variance effect due to ZO gradient estimator, we further introduce another RNN to learn the Gaussian sampling rule and dynamically guide the query direction sampling. Our learned optimizer outperforms hand-designed algorithms in terms of convergence rate and final solution on both synthetic and practical ZO optimization tasks (in particular, the black-box adversarial attack task, which is one of the most widely used tasks of ZO optimization). We finally conduct extensive analytical experiments to demonstrate the effectiveness of our proposed optimizer.

MetaPix: Few-Shot Video Retargeting

- Jessica Lee, Deva Ramanan, Rohit Girdhar
- abstract@open-review(Poster): We address the task of unsupervised retargeting of human actions from one video to another. We consider the challenging setting where only a few frames of the target is available. The core of our approach is a conditional generative model that can transcode input skeletal poses (automatically extracted with an off-the-shelf pose estimator) to output target frames. However, it is challenging to build a universal transcoder because humans can appear wildly different due to clothing and background scene geometry. Instead, we learn to adapt – or personalize – a universal generator to the particular human and background in the target. To do so, we make use of meta-learning to discover effective strategies for on-the-fly personalization. One significant benefit of meta-learning is that the personalized transcoder naturally enforces temporal coherence across its generated frames; all frames contain consistent clothing and background geometry of the target. We experiment on in-the-wild internet videos and images and show our approach improves over widely-used baselines for the task.

SlowMo: Improving Communication-Efficient Distributed SGD with Slow Momentum

- Jianyu Wang, Vinayak Tantia, Nicolas Ballas, Michael Rabbat
- abstract@open-review(Poster): Distributed optimization is essential for training large models on large datasets. Multiple approaches have been proposed to reduce the communication overhead in distributed training, such as synchronizing only after performing multiple local SGD steps, and decentralized methods (e.g., using gossip algorithms) to decouple communications among workers. Although these methods run faster than AllReduce-based methods, which use blocking communication before every update, the resulting models may be less accurate after the same number of updates. Inspired by the BMUF method of Chen & Huo (2016), we propose a slow momentum (SlowMo) framework, where workers periodically synchronize and perform a momentum update, after multiple iterations of a base optimization algorithm. Experiments on image classification and machine translation tasks demonstrate that SlowMo consistently yields improvements in optimization and generalization performance relative to the base optimizer, even when the

additional overhead is amortized over many updates so that the SlowMo runtime is on par with that of the base optimizer. We provide theoretical convergence guarantees showing that SlowMo converges to a stationary point of smooth non-convex losses. Since BMUF can be expressed through the SlowMo framework, our results also correspond to the first theoretical convergence guarantees for BMUF.

Pseudo-LiDAR++: Accurate Depth for 3D Object Detection in Autonomous Driving

- Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, Kilian Q. Weinberger
- abstract@open-review(Poster): Detecting objects such as cars and pedestrians in 3D plays an indispensable role in autonomous driving. Existing approaches largely rely on expensive LiDAR sensors for accurate depth information. While recently pseudo-LiDAR has been introduced as a promising alternative, at a much lower cost based solely on stereo images, there is still a notable performance gap. In this paper we provide substantial advances to the pseudo-LiDAR framework through improvements in stereo depth estimation. Concretely, we adapt the stereo network architecture and loss function to be more aligned with accurate depth estimation of faraway objects --- currently the primary weakness of pseudo-LiDAR. Further, we explore the idea to leverage cheaper but extremely sparse LiDAR sensors, which alone provide insufficient information for 3D detection, to de-bias our depth estimation. We propose a depth-propagation algorithm, guided by the initial depth estimates, to diffuse these few exact measurements across the entire depth map. We show on the KITTI object detection benchmark that our combined approach yields substantial improvements in depth estimation and stereo-based 3D object detection --- outperforming the previous state-of-the-art detection accuracy for faraway objects by 40%. Our code is available at https://github.com/mileyan/Pseudo_Lidar_V2.

Four Things Everyone Should Know to Improve Batch Normalization

- Cecilia Summers, Michael J. Dinneen
- abstract@open-review(Poster): A key component of most neural network architectures is the use of normalization layers, such as Batch Normalization. Despite its common use and large utility in optimizing deep architectures, it has been challenging both to generically improve upon Batch Normalization and to understand the circumstances that lend themselves to other enhancements. In this paper, we identify four improvements to the generic form of Batch Normalization and the circumstances under which they work, yielding performance gains across all batch sizes while requiring no additional computation during training. These contributions include proposing a method for reasoning about the current example in inference normalization statistics, fixing a training vs. inference discrepancy; recognizing and validating the powerful regularization effect of Ghost Batch Normalization for small and medium batch sizes; examining the effect of weight decay regularization on the scaling and shifting parameters γ and β; and identifying a new normalization algorithm for very small batch sizes by combining the strengths of Batch and Group Normalization. We validate our results empirically on six datasets: CIFAR-100, SVHN, Caltech-256, Oxford Flowers-102, CUB-2011, and ImageNet.

Learning to solve the credit assignment problem

- Benjamin James Lansdell, Prashanth Ravi Prakash, Konrad Paul Kording
- abstract@open-review(Poster): Backpropagation is driving today's artificial neural networks (ANNs). However, despite extensive research, it remains unclear if the brain implements this algorithm. Among neuroscientists, reinforcement learning (RL) algorithms are often seen as a realistic alternative: neurons can randomly introduce change, and use unspecific feedback signals to observe their effect on the cost and thus approximate their gradient. However, the convergence rate of such learning scales poorly with the number of involved neurons. Here we propose a hybrid learning approach. Each neuron uses an RL-type strategy to learn how to approximate the gradients that backpropagation would provide. We provide proof that our approach converges to the true gradient for certain classes of networks. In both feedforward and convolutional networks, we empirically show that our approach learns to approximate the gradient, and can match the performance of gradient-based learning. Learning feedback weights provides a biologically plausible mechanism of achieving good performance, without the need for precise, pre-specified learning rules.

Sampling-Free Learning of Bayesian Quantized Neural Networks

- Jiahao Su, Milan Cvitkovic, Furong Huang
- abstract@open-review(Poster): Bayesian learning of model parameters in neural networks is important in scenarios where estimates with well-calibrated uncertainty are important. In this paper, we propose Bayesian quantized networks (BQNs), quantized neural networks (QNNs) for which we learn a posterior distribution over their discrete parameters. We provide a set of efficient algorithms for learning and prediction in BQNs without the need to sample from their parameters or activations, which not only allows for differentiable learning in quantized models but also reduces the variance in gradients estimation. We evaluate BQNs on MNIST, Fashion-MNIST and KMNIST classification datasets compared against bootstrap ensemble of QNNs (E-QNN). We demonstrate BQNs achieve both lower predictive errors and better-calibrated uncertainties than E-QNN (with less than 20% of the negative log-likelihood).

DeFINE: Deep Factorized Input Token Embeddings for Neural Sequence Modeling

- Sachin Mehta, Rik Koncel-Kedziorski, Mohammad Rastegari, Hannaneh Hajishirzi
- abstract@open-review(Poster): For sequence models with large vocabularies, a majority of network parameters lie in the input and output layers. In this work, we describe a new method, DeFINE, for learning deep token representations efficiently. Our architecture uses a hierarchical structure with novel skip-connections which allows for the use of low dimensional input and output layers, reducing total parameters and training time while delivering similar or better performance versus existing methods. DeFINE can be incorporated easily in new or existing sequence models. Compared to state-of-the-art methods including adaptive input representations, this technique results in a 6% to 20% drop in perplexity. On WikiText-103, DeFINE reduces the total parameters of Transformer-XL by half with minimal impact on performance. On the Penn Treebank, DeFINE improves AWD-LSTM by 4 points with a 17% reduction in parameters, achieving comparable performance to state-of-the-art methods with fewer parameters. For machine translation, DeFINE improves the efficiency of the Transformer model by about 1.4 times while delivering similar performance.

DBA: Distributed Backdoor Attacks against Federated Learning

- Chulin Xie, Keli Huang, Pin-Yu Chen, Bo Li
- abstract@open-review(Poster): Backdoor attacks aim to manipulate a subset of training data by injecting adversarial triggers such that machine learning models trained on the tampered dataset will make arbitrarily (targeted) incorrect prediction on the testset with the same trigger embedded. While federated learning (FL) is capable of aggregating information provided by different parties for training a better model, its distributed learning methodology and inherently heterogeneous data distribution across parties may bring new vulnerabilities. In addition to recent centralized backdoor attacks on FL where each party embeds the same global trigger during training, we propose the distributed backdoor attack (DBA) --- a novel threat assessment framework developed by fully exploiting the distributed nature of FL. DBA decomposes a global trigger pattern into separate local patterns and embed them into the training set of different adversarial parties respectively. Compared to standard centralized backdoors, we show that DBA is substantially more persistent and stealthy against FL on diverse datasets such as finance and image data. We conduct extensive experiments to show that the attack success rate of DBA is significantly higher than centralized backdoors under different settings. Moreover, we find that distributed attacks are indeed more insidious, as DBA can evade two state-of-the-art robust FL algorithms against centralized backdoors. We also provide explanations for the effectiveness of DBA via feature visual interpretation and feature importance ranking. To further explore the properties of DBA, we test the attack performance by varying different trigger factors, including local trigger variations (size, gap, and location), scaling factor in FL, data distribution, and poison ratio and interval. Our proposed DBA and thorough evaluation results shed lights on characterizing the robustness of FL.

Fast is better than free: Revisiting adversarial training

- Eric Wong, Leslie Rice, J. Zico Kolter
- abstract@open-review(Poster): Adversarial training, a method for learning robust deep networks, is typically assumed to be more expensive than traditional training due to the necessity of constructing adversarial examples via a first-order method like projected gradient decent (PGD). In this paper, we make the surprising discovery that it is possible to train empirically robust models using a much weaker and cheaper adversary, an approach that was previously believed to be ineffective, rendering the method no more costly than standard training in practice. Specifically, we show that adversarial training with the fast gradient sign method (FGSM), when combined with random initialization, is as effective as PGD-based training but has significantly lower cost. Furthermore we show that FGSM adversarial training can be further accelerated by using standard techniques for efficient training of deep networks, allowing us to learn a robust CIFAR10 classifier with 45% robust accuracy at epsilon=8/255 in 6 minutes, and a robust ImageNet classifier with 43% robust accuracy at epsilon=2/255 in 12 hours, in comparison to past work based on ``free'' adversarial training which took 10 and 50 hours to reach the same respective thresholds.

Thieves on Sesame Street! Model Extraction of BERT-based APIs

- Kalpesh Krishna, Gaurav Singh Tomar, Ankur P. Parikh, Nicolas Papernot, Mohit Iyyer
- abstract@open-review(Poster): We study the problem of model extraction in natural language processing, in which an adversary with only query access to a victim model attempts to reconstruct a local copy of that model. Assuming that both the adversary and victim model fine-tune a large pretrained language model such as BERT (Devlin et al., 2019), we show that the adversary does not need any real training data to successfully mount the attack. In fact, the attacker need not even use grammatical or semantically meaningful queries: we show that random sequences of words coupled with task-specific heuristics form effective queries for model extraction on a diverse set of NLP tasks, including natural language inference and question answering. Our work thus highlights an exploit only made feasible by the shift towards transfer learning methods within the NLP community: for a query budget of a few hundred dollars, an attacker can extract a model that performs only slightly worse than the victim model. Finally, we study two defense strategies against model extraction—membership classification and API watermarking—which while successful against some adversaries can also be circumvented by more clever ones.

Understanding Knowledge Distillation in Non-autoregressive Machine Translation

- Chunting Zhou, Jiatao Gu, Graham Neubig
- abstract@open-review(Poster): Non-autoregressive machine translation (NAT) systems predict a sequence of output tokens in parallel, achieving substantial improvements in generation speed compared to autoregressive models. Existing NAT models usually rely on the technique of knowledge distillation, which creates the training data from a pretrained autoregressive model for better performance. Knowledge distillation is empirically useful, leading to large gains in accuracy for NAT models, but the reason for this success has, as of yet, been unclear. In this paper, we first design systematic experiments to investigate why knowledge distillation is crucial to NAT training. We find that knowledge distillation can reduce the complexity of data sets and help NAT to model the variations in the output data. Furthermore, a strong correlation is observed between the capacity of an NAT model and the optimal complexity of the distilled data for the best translation quality. Based on these findings, we further propose several approaches that can alter the complexity of data sets to improve the performance of NAT models. We achieve the state-of-the-art performance for the NAT-based models, and close the gap with the autoregressive baseline on WMT14 En-De benchmark.

Locality and Compositionality in Zero-Shot Learning

- Tristan Sylvain, Linda Petrini, Devon Hjelm
- abstract@open-review(Poster): In this work we study locality and compositionality in the context of learning representations for Zero Shot Learning (ZSL). In order to well-isolate the importance of these properties in learned representations, we impose the additional constraint that, differently from most recent work in ZSL, no pre-training on different datasets (e.g. ImageNet) is performed. The results of our experiment show how locality, in terms of small parts of the input, and compositionality, i.e. how well can the learned representations be expressed as a function of a smaller vocabulary, are both deeply related to generalization and motivate the focus on more local-aware models in future research directions for representation learning.

Recurrent neural circuits for contour detection

- Drew Linsley, *Junkyung Kim*, Alekh Ashok, Thomas Serre
- abstract@open-review(Poster): We introduce a deep recurrent neural network architecture that approximates visual cortical circuits (Mély et al., 2018). We show that this architecture, which we refer to as the γ-net, learns to solve contour detection tasks with better sample efficiency than state-of-the-art feedforward networks, while also exhibiting a classic perceptual illusion, known as the orientation-tilt illusion. Correcting this illusion significantly reduces \gnetw contour detection accuracy by driving it to prefer low-level edges over high-level object boundary contours. Overall, our study suggests that the orientation-tilt illusion is a byproduct of neural circuits that help biological visual systems achieve robust and efficient contour detection, and that incorporating these circuits in artificial neural networks can improve computer vision.

Disentangling neural mechanisms for perceptual grouping

- Junkyung Kim, *Drew Linsley*, Kalpit Thakkar, Thomas Serre
- abstract@open-review(Spotlight): Forming perceptual groups and individuating objects in visual scenes is an essential step towards visual intelligence. This ability is thought to arise in the brain from computations implemented by bottom-up, horizontal, and top-down connections between neurons. However, the relative contributions of these connections to perceptual grouping are poorly understood. We address this question by systematically evaluating neural network architectures featuring combinations bottom-up, horizontal, and top-down connections on two synthetic visual tasks, which stress low-level "Gestalt" vs. high-level object cues for perceptual grouping. We show that increasing the difficulty of either task strains learning for networks that rely solely on bottom-up connections. Horizontal connections resolve straining on tasks with Gestalt cues by supporting incremental grouping, whereas top-down connections rescue learning on tasks with high-level object cues by modifying coarse predictions about the position of the target object. Our findings dissociate the computational roles of bottom-up, horizontal and top-down connectivity, and demonstrate how a model featuring all of these interactions can more flexibly learn to form perceptual groups.

Chameleon: Adaptive Code Optimization for Expedited Deep Neural Network Compilation

- Byung Hoon Ahn, Prannoy Pilligundla, Amir Yazdanbakhsh, Hadi Esmaeilzadeh
- abstract@open-review(Poster): Achieving faster execution with shorter compilation time can foster further diversity and innovation in neural networks. However, the current paradigm of executing neural networks either relies on hand-optimized libraries, traditional compilation heuristics, or very recently genetic algorithms and other stochastic methods. These methods suffer from frequent costly hardware measurements rendering them not only too time consuming but also suboptimal. As such, we devise a solution that can learn to quickly adapt to a previously unseen design space for code optimization, both accelerating the search and improving the output performance. This solution dubbed Chameleon leverages reinforcement learning whose solution takes fewer steps to converge, and develops an adaptive sampling algorithm that not only focuses on the costly samples (real hardware measurements) on

representative points but also uses a domain-knowledge inspired logic to improve the samples itself. Experimentation with real hardware shows that Chameleon provides 4.45x speed up in optimization time over AutoTVM, while also improving inference time of the modern deep networks by 5.6%.

Intrinsic Motivation for Encouraging Synergistic Behavior

- Rohan Chitnis, Shubham Tulsiani, Saurabh Gupta, Abhinav Gupta
- abstract@open-review(Poster): We study the role of intrinsic motivation as an exploration bias for reinforcement learning in sparse-reward synergistic tasks, which are tasks where multiple agents must work together to achieve a goal they could not individually. Our key idea is that a good guiding principle for intrinsic motivation in synergistic tasks is to take actions which affect the world in ways that would not be achieved if the agents were acting on their own. Thus, we propose to incentivize agents to take (joint) actions whose effects cannot be predicted via a composition of the predicted effect for each individual agent. We study two instantiations of this idea, one based on the true states encountered, and another based on a dynamics model trained concurrently with the policy. While the former is simpler, the latter has the benefit of being analytically differentiable with respect to the action taken. We validate our approach in robotic bimanual manipulation and multi-agent locomotion tasks with sparse rewards; we find that our approach yields more efficient learning than both 1) training with only the sparse reward and 2) using the typical surprise-based formulation of intrinsic motivation, which does not bias toward synergistic behavior. Videos are available on the project webpage: https://sites.google.com/view/iclr2020-synergistic.

RaCT: Toward Amortized Ranking-Critical Training For Collaborative Filtering

- Sam Lobel, *Chunyuan Li*, Jianfeng Gao, Lawrence Carin
- abstract@open-review(Poster): We investigate new methods for training collaborative filtering models based on actor-critic reinforcement learning, to more directly maximize ranking-based objective functions. Specifically, we train a critic network to approximate ranking-based metrics, and then update the actor network to directly optimize against the learned metrics. In contrast to traditional learning-to-rank methods that require re-running the optimization procedure for new lists, our critic-based method amortizes the scoring process with a neural network, and can directly provide the (approximate) ranking scores for new lists.

We demonstrate the actor-critic's ability to significantly improve the performance of a variety of prediction models, and achieve better or comparable performance to a variety of strong baselines on three large-scale datasets.

ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut
- abstract@open-review(Spotlight): Increasing model size when pretraining natural language representations often results in improved performance on downstream tasks. However, at some point further model increases become harder due to GPU/TPU memory limitations and longer training times. To address these problems, we present two parameter-reduction techniques to lower memory consumption and increase the training speed of BERT~\citep{devlin2018bert}. Comprehensive empirical evidence shows that our proposed methods lead to models that scale much better compared to the original BERT. We also use a self-supervised loss that focuses on modeling inter-sentence coherence, and show it consistently helps downstream tasks with multi-sentence inputs. As a result, our best model establishes new state-of-the-art results on the GLUE, RACE, and \squad benchmarks while having fewer parameters compared to BERT-large. The code and the pretrained models are available at https://github.com/google-research/ALBERT.

Sign-OPT: A Query-Efficient Hard-label Adversarial Attack

- Minhao Cheng, Simranjit Singh, Patrick H. Chen, Pin-Yu Chen, Sijia Liu, Cho-Jui Hsieh
- abstract@open-review(Poster): We study the most practical problem setup for evaluating adversarial robustness of a machine learning system with limited access: the hard-label black-box attack setting for generating adversarial examples, where limited model queries are allowed and only the decision is provided to a queried data input. Several algorithms have been proposed for this problem but they typically require huge amount (>20,000) of queries for attacking one example. Among them, one of the state-of-the-art approaches (Cheng et al., 2019) showed that hard-label attack can be modeled as an optimization problem where the objective function can be evaluated by binary search with additional model queries, thereby a zeroth order optimization algorithm can be applied. In this paper, we adopt the same optimization formulation but propose to directly estimate the sign of gradient at any direction instead of the gradient itself, which enjoys the benefit of single query. Using this single query oracle for retrieving sign of directional derivative, we develop a novel query-efficient Sign-OPT approach for hard-label black-box attack. We provide a convergence analysis of the new algorithm and conduct experiments on several models on MNIST, CIFAR-10 and ImageNet. We find that Sign-OPT attack consistently requires 5X to 10X fewer queries when compared to the current state-of-the-art approaches, and usually converges to an adversarial example with smaller perturbation.

Playing the lottery with rewards and multiple languages: lottery tickets in RL and NLP

- Haonan Yu, Sergey Edunov, Yuandong Tian, Ari S. Morcos
- abstract@open-review(Poster): The lottery ticket hypothesis proposes that over-parameterization of deep neural networks (DNNs) aids training by increasing the probability of a “lucky” sub-network initialization being present rather than by helping the optimization process (Frankle& Carbin, 2019). Intriguingly, this phenomenon suggests that initialization strategies for DNNs can be improved substantially, but the lottery ticket hypothesis has only previously been tested in the context of supervised learning for natural image tasks. Here, we evaluate whether “winning ticket” initializations exist in two different domains: natural language processing (NLP) and reinforcement learning (RL).For NLP, we examined both recurrent LSTM models and large-scale Transformer models (Vaswani et al., 2017). For RL, we analyzed a number of discrete-action space tasks, including both classic control and pixel control. Consistent with workin supervised image classification, we confirm that winning ticket initializations generally outperform parameter-matched random initializations, even at extreme pruning rates for both NLP and RL. Notably, we are able to find winning ticket initializations for Transformers which enable models one-third the size to achieve nearly equivalent performance. Together, these results suggest that the lottery ticket hypothesis is not restricted to supervised learning of natural images, but rather represents a broader phenomenon in DNNs.

Learning Space Partitions for Nearest Neighbor Search

- Yihe Dong, Piotr Indyk, Ilya Razenshteyn, Tal Wagner
- abstract@open-review(Poster): Space partitions of $\mathbb{R}^d$ underlie a vast and important class of fast nearest neighbor search (NNS) algorithms. Inspired by recent theoretical work on NNS for general metric spaces (Andoni et al. 2018b,c), we develop a new framework for building space partitions reducing the problem to balanced graph partitioning followed by supervised classification. We instantiate this general approach with the KaHIP graph partitioner (Sanders and Schulz 2013) and neural networks, respectively, to obtain a new partitioning procedure called Neural Locality-Sensitive Hashing (Neural LSH). On several standard benchmarks for NNS (Aumuller et al. 2017), our experiments show that the partitions obtained by Neural LSH consistently outperform partitions found by quantization-based and tree-based methods as well as classic, data-oblivious LSH.

DeepV2D: Video to Depth with Differentiable Structure from Motion

- Zachary Teed, Jia Deng
- abstract@open-review(Poster): We propose DeepV2D, an end-to-end deep learning architecture for predicting depth from video. DeepV2D combines the representation ability of neural networks with the geometric principles governing image formation. We compose a collection of classical geometric

algorithms, which are converted into trainable modules and combined into an end-to-end differentiable architecture. DeepV2D interleaves two stages: motion estimation and depth estimation. During inference, motion and depth estimation are alternated and converge to accurate depth.

The Break-Even Point on Optimization Trajectories of Deep Neural Networks

- Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, *Krzysztof Geras*
- abstract@open-review(Spotlight): The early phase of training of deep neural networks is critical for their final performance. In this work, we study how the hyperparameters of stochastic gradient descent (SGD) used in the early phase of training affect the rest of the optimization trajectory. We argue for the existence of the "``break-even" point on this trajectory, beyond which the curvature of the loss surface and noise in the gradient are implicitly regularized by SGD. In particular, we demonstrate on multiple classification tasks that using a large learning rate in the initial phase of training reduces the variance of the gradient, and improves the conditioning of the covariance of gradients. These effects are beneficial from the optimization perspective and become visible after the break-even point. Complementing prior work, we also show that using a low learning rate results in bad conditioning of the loss surface even for a neural network with batch normalization layers. In short, our work shows that key properties of the loss surface are strongly influenced by SGD in the early phase of training. We argue that studying the impact of the identified effects on generalization is a promising future direction.

Towards Better Understanding of Adaptive Gradient Algorithms in Generative Adversarial Nets

- Mingrui Liu, Youssef Mroueh, Jerret Ross, Wei Zhang, Xiaodong Cui, Payel Das, Tianbao Yang
- abstract@open-review(Poster): Adaptive gradient algorithms perform gradient-based updates using the history of gradients and are ubiquitous in training deep neural networks. While adaptive gradient methods theory is well understood for minimization problems, the underlying factors driving their empirical success in min-max problems such as GANs remain unclear. In this paper, we aim at bridging this gap from both theoretical and empirical perspectives. First, we analyze a variant of Optimistic Stochastic Gradient (OSG) proposed in~\citep{daskalakis2017training} for solving a class of non-convex non-concave min-max problem and establish $O(\epsilon^{-4})$ complexity for finding ϵ-first-order stationary point, in which the algorithm only requires invoking one stochastic first-order oracle while enjoying state-of-the-art iteration complexity achieved by stochastic extragradient method by~\citep{iusem2017extragradient}. Then we propose an adaptive variant of OSG named Optimistic Adagrad (OAdagrad) and reveal an \emph{improved} adaptive complexity $O\left(\epsilon^{-\frac{2}{1-\alpha}}\right)$, where α characterizes the growth rate of the cumulative stochastic gradient and $0\leq \alpha\leq 1/2$. To the best of our knowledge, this is the first work for establishing adaptive complexity in non-convex non-concave min-max optimization. Empirically, our experiments show that indeed adaptive gradient algorithms outperform their non-adaptive counterparts in GAN training. Moreover, this observation can be explained by the slow growth rate of the cumulative stochastic gradient, as observed empirically.

Robust And Interpretable Blind Image Denoising Via Bias-Free Convolutional Neural Networks

- Sreyas Mohan, Zahra Kadkhodaie, Eero P. Simoncelli, Carlos Fernandez-Granda
- abstract@open-review(Poster): We study the generalization properties of deep convolutional neural networks for image denoising in the presence of varying noise levels. We provide extensive empirical evidence that current state-of-the-art architectures systematically overfit to the noise levels in the training set, performing very poorly at new noise levels. We show that strong generalization can be achieved through a simple architectural modification: removing all additive constants. The resulting "bias-free" networks attain state-of-the-art performance over a broad range of noise levels, even when trained over a limited range. They are also locally linear, which enables direct analysis with linear-algebraic tools. We show that the denoising map can be visualized locally as a filter that adapts to both image structure and noise level. In addition, our analysis reveals that deep networks implicitly perform a projection onto an adaptively-selected low-dimensional subspace, with dimensionality inversely proportional to noise level, that captures features of natural images.

CM3: Cooperative Multi-goal Multi-stage Multi-agent Reinforcement Learning

- Jiachen Yang, Alireza Nakhaei, David Isele, Kikuo Fujimura, Hongyuan Zha
- abstract@open-review(Poster): A variety of cooperative multi-agent control problems require agents to achieve individual goals while contributing to collective success. This multi-goal multi-agent setting poses difficulties for recent algorithms, which primarily target settings with a single global reward, due to two new challenges: efficient exploration for learning both individual goal attainment and cooperation for others' success, and credit-assignment for interactions between actions and goals of different agents. To address both challenges, we restructure the problem into a novel two-stage curriculum, in which single-agent goal attainment is learned prior to learning multi-agent cooperation, and we derive a new multi-goal multi-agent policy gradient with a credit function for localized credit assignment. We use a function augmentation scheme to bridge value and policy functions across the curriculum. The complete architecture, called CM3, learns significantly faster than direct adaptations of existing algorithms on three challenging multi-goal multi-agent problems: cooperative navigation in difficult formations, negotiating multi-vehicle lane changes in the SUMO traffic simulator, and strategic cooperation in a Checkers environment.

Deep Imitative Models for Flexible Inference, Planning, and Control

- Nicholas Rhinehart, Rowan McAllister, Sergey Levine
- abstract@open-review(Poster): Imitation Learning (IL) is an appealing approach to learn desirable autonomous behavior. However, directing IL to achieve arbitrary goals is difficult. In contrast, planning-based algorithms use dynamics models and reward functions to achieve goals. Yet, reward functions that evoke desirable behavior are often difficult to specify. In this paper, we propose "Imitative Models" to combine the benefits of IL and goal-directed planning. Imitative Models are probabilistic predictive models of desirable behavior able to plan interpretable expert-like trajectories to achieve specified goals. We derive families of flexible goal objectives, including constrained goal regions, unconstrained goal sets, and energy-based goals. We show that our method can use these objectives to successfully direct behavior. Our method substantially outperforms six IL approaches and a planning-based approach in a dynamic simulated autonomous driving task, and is efficiently learned from expert demonstrations without online data collection. We also show our approach is robust to poorly-specified goals, such as goals on the wrong side of the road.

The Logical Expressiveness of Graph Neural Networks

- Pablo Barceló, Egor V. Kostylev, Mikael Monet, Jorge Pérez, Juan Reutter, Juan Pablo Silva
- abstract@open-review(Spotlight): The ability of graph neural networks (GNNs) for distinguishing nodes in graphs has been recently characterized in terms of the Weisfeiler-Lehman (WL) test for checking graph isomorphism. This characterization, however, does not settle the issue of which Boolean node classifiers (i.e., functions classifying nodes in graphs as true or false) can be expressed by GNNs. We tackle this problem by focusing on Boolean classifiers expressible as formulas in the logic FOC2, a well-studied fragment of first order logic. FOC2 is tightly related to the WL test, and hence to GNNs. We start by studying a popular class of GNNs, which we call AC-GNNs, in which the features of each node in the graph are updated, in successive layers, only in terms of the features of its neighbors. We show that this class of GNNs is too weak to capture all FOC2 classifiers, and provide a syntactic characterization of the largest subclass of FOC2 classifiers that can be captured by AC-GNNs. This subclass coincides with a logic heavily used by the knowledge representation community. We then look at what needs to be added to AC-GNNs for capturing all FOC2 classifiers. We show that it suffices to add readout functions, which allow to update the features of a node not only in terms of its neighbors, but also in terms of a global attribute vector. We call GNNs of this kind ACR-GNNs. We experimentally validate our findings showing that, on synthetic data conforming to FOC2 formulas, AC-GNNs struggle to fit the training data while ACR-GNNs can generalize even to graphs of sizes not seen during training.

Pre-training Tasks for Embedding-based Large-scale Retrieval

- Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, Sanjiv Kumar
- abstract@open-review(Poster): We consider the large-scale query-document retrieval problem: given a query (e.g., a question), return the set of relevant documents (e.g., paragraphs containing the answer) from a large document corpus. This problem is often solved in two steps. The retrieval phase first reduces the solution space, returning a subset of candidate documents. The scoring phase then re-ranks the documents. Critically, the retrieval algorithm not only desires high recall but also requires to be highly efficient, returning candidates in time sublinear to the number of documents. Unlike the scoring phase witnessing significant advances recently due to the BERT-style pre-training tasks on cross-attention models, the retrieval phase remains less well studied. Most previous works rely on classic Information Retrieval (IR) methods such as BM-25 (token matching + TF-IDF weights). These models only accept sparse handcrafted features and can not be optimized for different downstream tasks of interest. In this paper, we conduct a comprehensive study on the embedding-based retrieval models. We show that the key ingredient of learning a strong embedding-based Transformer model is the set of pre-training tasks. With adequately designed paragraph-level pre-training tasks, the Transformer models can remarkably improve over the widely-used BM-25 as well as embedding models without Transformers. The paragraph-level pre-training tasks we studied are Inverse Cloze Task (ICT), Body First Selection (BFS), Wiki Link Prediction (WLP), and the combination of all three.

Are Transformers universal approximators of sequence-to-sequence functions?

- Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, Sanjiv Kumar
- abstract@open-review(Poster): Despite the widespread adoption of Transformer models for NLP tasks, the expressive power of these models is not well-understood. In this paper, we establish that Transformer models are universal approximators of continuous permutation equivariant sequence-to-sequence functions with compact support, which is quite surprising given the amount of shared parameters in these models. Furthermore, using positional encodings, we circumvent the restriction of permutation equivariance, and show that Transformer models can universally approximate arbitrary continuous sequence-to-sequence functions on a compact domain. Interestingly, our proof techniques clearly highlight the different roles of the self-attention and the feed-forward layers in Transformers. In particular, we prove that fixed width self-attention layers can compute contextual mappings of the input sequences, playing a key role in the universal approximation property of Transformers. Based on this insight from our analysis, we consider other simpler alternatives to self-attention layers and empirically evaluate them.

Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples

- Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, Hugo Larochelle
- abstract@open-review(Poster): Few-shot classification refers to learning a classifier for new classes given only a few examples. While a plethora of models have emerged to tackle it, we find the procedure and datasets that are used to assess their progress lacking. To address this limitation, we propose Meta-Dataset: a new benchmark for training and evaluating models that is large-scale, consists of diverse datasets, and presents more realistic tasks. We experiment with popular baselines and meta-learners on Meta-Dataset, along with a competitive method that we propose. We analyze performance as a function of various characteristics of test tasks and examine the models' ability to leverage diverse training sources for improving their generalization. We also propose a new set of baselines for quantifying the benefit of meta-learning in Meta-Dataset. Our extensive experimentation has uncovered important research challenges and we hope to inspire work in these directions.

One-Shot Pruning of Recurrent Neural Networks by Jacobian Spectrum Evaluation

- Shunshi Zhang, Bradly C. Stadie
- abstract@open-review(Poster): Recent advances in the sparse neural network literature have made it possible to prune many large feed forward and convolutional networks with only a small quantity of data. Yet, these same techniques often falter when applied to the problem of recovering sparse recurrent networks. These failures are quantitative: when pruned with recent techniques, RNNs typically obtain worse performance than they do under a simple random pruning scheme. The failures are also qualitative: the distribution of active weights in a pruned LSTM or GRU network tend to be concentrated in specific neurons and gates, and not well dispersed across the entire architecture. We seek to rectify both the quantitative and qualitative issues with recurrent network pruning by introducing a new recurrent pruning objective derived from the spectrum of the recurrent Jacobian. Our objective is data efficient (requiring only 64 data points to prune the network), easy to implement, and produces 95 % sparse GRUs that significantly improve on existing baselines. We evaluate on sequential MNIST, Billion Words, and Wikitext.

Differentially Private Meta-Learning

- Jeffrey Li, Mikhail Khodak, Sebastian Caldas, Ameet Talwalkar
- abstract@open-review(Poster): Parameter-transfer is a well-known and versatile approach for meta-learning, with applications including few-shot learning, federated learning, with personalization, and reinforcement learning. However, parameter-transfer algorithms often require sharing models that have been trained on the samples from specific tasks, thus leaving the task-owners susceptible to breaches of privacy. We conduct the first formal study of privacy in this setting and formalize the notion of task-global differential privacy as a practical relaxation of more commonly studied threat models. We then propose a new differentially private algorithm for gradient-based parameter transfer that not only satisfies this privacy requirement but also retains provable transfer learning guarantees in convex settings. Empirically, we apply our analysis to the problems of federated learning with personalization and few-shot classification, showing that allowing the relaxation to task-global privacy from the more commonly studied notion of local privacy leads to dramatically increased performance in recurrent neural language modeling and image classification.

CLEVRER: Collision Events for Video Representation and Reasoning

- Kexin Yi, *Chuang Gan*, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, Joshua B. Tenenbaum
- abstract@open-review(Spotlight): The ability to reason about temporal and causal events from videos lies at the core of human intelligence. Most video reasoning benchmarks, however, focus on pattern recognition from complex visual and language input, instead of on causal structure. We study the complementary problem, exploring the temporal and causal structures behind videos of objects with simple visual appearance. To this end, we introduce the CoLlision Events for Video REpresentation and Reasoning (CLEVRER) dataset, a diagnostic video dataset for systematic evaluation of computational models on a wide range of reasoning tasks. Motivated by the theory of human casual judgment, CLEVRER includes four types of question: descriptive (e.g., 'what color'), explanatory ('what's responsible for'), predictive ('what will happen next'), and counterfactual ('what if'). We evaluate various state-of-the-art models for visual reasoning on our benchmark. While these models thrive on the perception-based task (descriptive), they perform poorly on the causal tasks (explanatory, predictive and counterfactual), suggesting that a principled approach for causal reasoning should incorporate the capability of both perceiving complex visual and language inputs, and understanding the underlying dynamics and causal relations. We also study an oracle model that explicitly combines these components via symbolic representations.

Learning Compositional Koopman Operators for Model-Based Control

- Yunzhu Li, Hao He, Jiajun Wu, Dina Katabi, Antonio Torralba

- abstract@open-review(Spotlight): Finding an embedding space for a linear approximation of a nonlinear dynamical system enables efficient system identification and control synthesis. The Koopman operator theory lays the foundation for identifying the nonlinear-to-linear coordinate transformations with data-driven methods. Recently, researchers have proposed to use deep neural networks as a more expressive class of basis functions for calculating the Koopman operators. These approaches, however, assume a fixed dimensional state space; they are therefore not applicable to scenarios with a variable number of objects. In this paper, we propose to learn compositional Koopman operators, using graph neural networks to encode the state into object-centric embeddings and using a block-wise linear transition matrix to regularize the shared structure across objects. The learned dynamics can quickly adapt to new environments of unknown physical parameters and produce control signals to achieve a specified goal. Our experiments on manipulating ropes and controlling soft robots show that the proposed method has better efficiency and generalization ability than existing baselines.

Bridging Mode Connectivity in Loss Landscapes and Adversarial Robustness

- Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, Xue Lin
- abstract@open-review(Poster): Mode connectivity provides novel geometric insights on analyzing loss landscapes and enables building high-accuracy pathways between well-trained neural networks. In this work, we propose to employ mode connectivity in loss landscapes to study the adversarial robustness of deep neural networks, and provide novel methods for improving this robustness. Our experiments cover various types of adversarial attacks applied to different network architectures and datasets. When network models are tampered with backdoor or error-injection attacks, our results demonstrate that the path connection learned using limited amount of bonafide data can effectively mitigate adversarial effects while maintaining the original accuracy on clean data. Therefore, mode connectivity provides users with the power to repair backdoored or error-injected models. We also use mode connectivity to investigate the loss landscapes of regular and robust models against evasion attacks. Experiments show that there exists a barrier in adversarial robustness loss on the path connecting regular and adversarially-trained models. A high correlation is observed between the adversarial robustness loss and the largest eigenvalue of the input Hessian matrix, for which theoretical justifications are provided. Our results suggest that mode connectivity offers a holistic tool and practical means for evaluating and improving adversarial robustness.

Overlearning Reveals Sensitive Attributes

- Congzheng Song, Vitaly Shmatikov
- abstract@open-review(Poster): ``"Overlearning" means that a model trained for a seemingly simple objective implicitly learns to recognize attributes and concepts that are (1) not part of the learning objective, and (2) sensitive from a privacy or bias perspective. For example, a binary gender classifier of facial images also learns to recognize races, even races that are not represented in the training data, and identities.

We demonstrate overlearning in several vision and NLP models and analyze its harmful consequences. First, inference-time representations of an overlearned model reveal sensitive attributes of the input, breaking privacy protections such as model partitioning. Second, an overlearned model can be "`re-purposed" for a different, privacy-violating task even in the absence of the original training data.

We show that overlearning is intrinsic for some tasks and cannot be prevented by censoring unwanted attributes. Finally, we investigate where, when, and why overlearning happens during model training.

Optimal Strategies Against Generative Attacks

- Roy Mor, Erez Peterfreund, Matan Gavish, Amir Globerson
- abstract@open-review(Talk): Generative neural models have improved dramatically recently. With this progress comes the risk that such models will be used to attack systems that rely on sensor data for authentication and anomaly detection. Many such learning systems are installed worldwide, protecting critical infrastructure or private data against malfunction and cyber attacks. We formulate the scenario of such an authentication system facing generative impersonation attacks, characterize it from a theoretical perspective and explore its practical implications. In particular, we ask fundamental theoretical questions in learning, statistics and information theory: How hard is it to detect a "fake reality"? How much data does the attacker need to collect before it can reliably generate nominally-looking artificial data? Are there optimal strategies for the attacker or the authenticator? We cast the problem as a maximin game, characterize the optimal strategy for both attacker and authenticator in the general case, and provide the optimal strategies in closed form for the case of Gaussian source distributions. Our analysis reveals the structure of the optimal attack and the relative importance of data collection for both authenticator and attacker. Based on these insights we design practical learning approaches and show that they result in models that are more robust to various attacks on real-world data.

Adversarially robust transfer learning

- Ali Shafahi, Parsa Saadatpanah, Chen Zhu, Amin Ghiasi, Christoph Studer, David Jacobs, Tom Goldstein
- abstract@open-review(Poster): Transfer learning, in which a network is trained on one task and re-purposed on another, is often used to produce neural network classifiers when data is scarce or full-scale training is too costly. When the goal is to produce a model that is not only accurate but also adversarially robust, data scarcity and computational limitations become even more cumbersome. We consider robust transfer learning, in which we transfer not only performance but also robustness from a source model to a target domain. We start by observing that robust networks contain robust feature extractors. By training classifiers on top of these feature extractors, we produce new models that inherit the robustness of their parent networks. We then consider the case of "fine tuning" a network by re-training end-to-end in the target domain. When using lifelong learning strategies, this process preserves the robustness of the source network while achieving high accuracy. By using such strategies, it is possible to produce accurate and robust models with little data, and without the cost of adversarial training. Additionally, we can improve the generalization of adversarially trained models, while maintaining their robustness.

Doubly Robust Bias Reduction in Infinite Horizon Off-Policy Estimation

- Ziyang Tang, *Yihao Feng*, Lihong Li, Dengyong Zhou, Qiang Liu
- abstract@open-review(Spotlight): Infinite horizon off-policy policy evaluation is a highly challenging task due to the excessively large variance of typical importance sampling (IS) estimators. Recently, Liu et al. (2018) proposed an approach that significantly reduces the variance of infinite-horizon off-policy evaluation by estimating the stationary density ratio, but at the cost of introducing potentially high risks due to the error in density ratio estimation. In this paper, we develop a bias-reduced augmentation of their method, which can take advantage of a learned value function to obtain higher accuracy. Our method is doubly robust in that the bias vanishes when either the density ratio or value function estimation is perfect. In general, when either of them is accurate, the bias can also be reduced. Both theoretical and empirical results show that our method yields significant advantages over previous methods.

Learning to Link

- Maria-Florina Balcan, Travis Dick, Manuel Lang
- abstract@open-review(Poster): Clustering is an important part of many modern data analysis pipelines, including network analysis and data retrieval. There are many different clustering algorithms developed by various communities, and it is often not clear which algorithm will give the best performance on a specific clustering task. Similarly, we often have multiple ways to measure distances between data points, and the best clustering performance might require a non-trivial combination of those metrics. In this work, we study data-driven algorithm selection and metric learning for clustering problems, where the goal is to simultaneously learn the best algorithm and metric for a specific application. The family of clustering algorithms we consider is

parameterized linkage based procedures that includes single and complete linkage. The family of distance functions we learn over are convex combinations of base distance functions. We design efficient learning algorithms which receive samples from an application-specific distribution over clustering instances and learn a near-optimal distance and clustering algorithm from these classes. We also carry out a comprehensive empirical evaluation of our techniques showing that they can lead to significantly improved clustering performance on real-world datasets.

Detecting Extrapolation with Local Ensembles

- David Madras, James Atwood, Alexander D'Amour
- abstract@open-review(Poster): We present local ensembles, a method for detecting extrapolation at test time in a pre-trained model. We focus on underdetermination as a key component of extrapolation: we aim to detect when many possible predictions are consistent with the training data and model class. Our method uses local second-order information to approximate the variance of predictions across an ensemble of models from the same class. We compute this approximation by estimating the norm of the component of a test point's gradient that aligns with the low-curvature directions of the Hessian, and provide a tractable method for estimating this quantity. Experimentally, we show that our method is capable of detecting when a pre-trained model is extrapolating on test data, with applications to out-of-distribution detection, detecting spurious correlates, and active learning.

Global Relational Models of Source Code

- Vincent J. Hellendoorn, Charles Sutton, Rishabh Singh, Petros Maniatis, David Bieber
- abstract@open-review(Poster): Models of code can learn distributed representations of a program's syntax and semantics to predict many non-trivial properties of a program. Recent state-of-the-art models leverage highly structured representations of programs, such as trees, graphs and paths therein (e.g. data-flow relations), which are precise and abundantly available for code. This provides a strong inductive bias towards semantically meaningful relations, yielding more generalizable representations than classical sequence-based models. Unfortunately, these models primarily rely on graph-based message passing to represent relations in code, which makes them de facto local due to the high cost of message-passing steps, quite in contrast to modern, global sequence-based models, such as the Transformer. In this work, we bridge this divide between global and structured models by introducing two new hybrid model families that are both global and incorporate structural bias: Graph Sandwiches, which wrap traditional (gated) graph message-passing layers in sequential message-passing layers; and Graph Relational Embedding Attention Transformers (GREAT for short), which bias traditional Transformers with relational information from graph edge types. By studying a popular, non-trivial program repair task, variable-misuse identification, we explore the relative merits of traditional and hybrid model families for code representation. Starting with a graph-based model that already improves upon the prior state-of-the-art for this task by 20%, we show that our proposed hybrid models improve an additional 10-15%, while training both faster and using fewer parameters.

Selection via Proxy: Efficient Data Selection for Deep Learning

- Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, Matei Zaharia
- abstract@open-review(Poster): Data selection methods, such as active learning and core-set selection, are useful tools for machine learning on large datasets. However, they can be prohibitively expensive to apply in deep learning because they depend on feature representations that need to be learned. In this work, we show that we can greatly improve the computational efficiency by using a small proxy model to perform data selection (e.g., selecting data points to label for active learning). By removing hidden layers from the target model, using smaller architectures, and training for fewer epochs, we create proxies that are an order of magnitude faster to train. Although these small proxy models have higher error rates, we find that they empirically provide useful signals for data selection. We evaluate this "selection via proxy" (SVP) approach on several data selection tasks across five datasets: CIFAR10, CIFAR100, ImageNet, Amazon Review Polarity, and Amazon Review Full. For active learning, applying SVP can give an order of magnitude improvement in data selection runtime (i.e., the time it takes to repeatedly train and select points) without significantly increasing the final error (often within 0.1%). For core-set selection on CIFAR10, proxies that are over 10× faster to train than their larger, more accurate targets can remove up to 50% of the data without harming the final accuracy of the target, leading to a 1.6× end-to-end training time improvement.

Short and Sparse Deconvolution --- A Geometric Approach

- Yenson Lau, Qing Qu, Han-Wen Kuo, Pengcheng Zhou, Yuqian Zhang, John Wright
- abstract@open-review(Poster): Short-and-sparse deconvolution (SaSD) is the problem of extracting localized, recurring motifs in signals with spatial or temporal structure. Variants of this problem arise in applications such as image deblurring, microscopy, neural spike sorting, and more. The problem is challenging in both theory and practice, as natural optimization formulations are nonconvex. Moreover, practical deconvolution problems involve smooth motifs (kernels) whose spectra decay rapidly, resulting in poor conditioning and numerical challenges. This paper is motivated by recent theoretical advances \citep{zhang2017global,kuo2019geometry}, which characterize the optimization landscape of a particular nonconvex formulation of SaSD. This is used to derive a provable algorithm that exactly solves certain non-practical instances of the SaSD problem. We leverage the key ideas from this theory (sphere constraints, data-driven initialization) to develop a practical algorithm, which performs well on data arising from a range of application areas. We highlight key additional challenges posed by the ill-conditioning of real SaSD problems and suggest heuristics (acceleration, continuation, reweighting) to mitigate them. Experiments demonstrate the performance and generality of the proposed method.

Stochastic Weight Averaging in Parallel: Large-Batch Training That Generalizes Well

- Vipul Gupta, Santiago Akle Serrano, Dennis DeCoste
- abstract@open-review(Poster): We propose Stochastic Weight Averaging in Parallel (SWAP), an algorithm to accelerate DNN training. Our algorithm uses large mini-batches to compute an approximate solution quickly and then refines it by averaging the weights of multiple models computed independently and in parallel. The resulting models generalize equally well as those trained with small mini-batches but are produced in a substantially shorter time. We demonstrate the reduction in training time and the good generalization performance of the resulting models on the computer vision datasets CIFAR10, CIFAR100, and ImageNet.

Adjustable Real-time Style Transfer

- Mohammad Babaeizadeh, Golnaz Ghiasi
- abstract@open-review(Poster): Artistic style transfer is the problem of synthesizing an image with content similar to a given image and style similar to another. Although recent feed-forward neural networks can generate stylized images in real-time, these models produce a single stylization given a pair of style/content images, and the user doesn't have control over the synthesized output. Moreover, the style transfer depends on the hyper-parameters of the model with varying ``optimum" for different input images. Therefore, if the stylized output is not appealing to the user, she/he has to try multiple models or retrain one with different hyper-parameters to get a favorite stylization. In this paper, we address these issues by proposing a novel method which allows adjustment of crucial hyper-parameters, after the training and in real-time, through a set of manually adjustable parameters. These parameters enable the user to modify the synthesized outputs from the same pair of style/content images, in search of a favorite stylized image. Our quantitative and qualitative experiments indicate how adjusting these parameters is comparable to retraining the model with different hyper-parameters. We also demonstrate how these parameters can be randomized to generate results which are diverse but still very similar in style and content.

Dynamics-Aware Unsupervised Discovery of Skills

- Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, Karol Hausman
- abstract@open-review(Talk): Conventionally, model-based reinforcement learning (MBRL) aims to learn a global model for the dynamics of the environment. A good model can potentially enable planning algorithms to generate a large variety of behaviors and solve diverse tasks. However, learning an accurate model for complex dynamical systems is difficult, and even then, the model might not generalize well outside the distribution of states on which it was trained. In this work, we combine model-based learning with model-free learning of primitives that make model-based planning easy. To that end, we aim to answer the question: how can we discover skills whose outcomes are easy to predict? We propose an unsupervised learning algorithm, Dynamics-Aware Discovery of Skills (DADS), which simultaneously discovers predictable behaviors and learns their dynamics. Our method can leverage continuous skill spaces, theoretically, allowing us to learn infinitely many behaviors even for high-dimensional state-spaces. We demonstrate that zero-shot planning in the learned latent space significantly outperforms standard MBRL and model-free goal-conditioned RL, can handle sparse-reward tasks, and substantially improves over prior hierarchical RL methods for unsupervised skill discovery.

Unpaired Point Cloud Completion on Real Scans using Adversarial Training

- Xuelin Chen, Baoquan Chen, Niloy J. Mitra
- abstract@open-review(Poster): As 3D scanning solutions become increasingly popular, several deep learning setups have been developed for the task of scan completion, i.e., plausibly filling in regions that were missed in the raw scans. These methods, however, largely rely on supervision in the form of paired training data, i.e., partial scans with corresponding desired completed scans. While these methods have been successfully demonstrated on synthetic data, the approaches cannot be directly used on real scans in absence of suitable paired training data. We develop a first approach that works directly on input point clouds, does not require paired training data, and hence can directly be applied to real scans for scan completion. We evaluate the approach qualitatively on several real-world datasets (ScanNet, Matterport3D, KITTI), quantitatively on 3D-EPN shape completion benchmark dataset, and demonstrate realistic completions under varying levels of incompleteness.

Efficient Riemannian Optimization on the Stiefel Manifold via the Cayley Transform

- Jun Li, Fuxin Li, Sinisa Todorovic
- abstract@open-review(Poster): Strictly enforcing orthonormality constraints on parameter matrices has been shown advantageous in deep learning. This amounts to Riemannian optimization on the Stiefel manifold, which, however, is computationally expensive. To address this challenge, we present two main contributions: (1) A new efficient retraction map based on an iterative Cayley transform for optimization updates, and (2) An implicit vector transport mechanism based on the combination of a projection of the momentum and the Cayley transform on the Stiefel manifold. We specify two new optimization algorithms: Cayley SGD with momentum, and Cayley ADAM on the Stiefel manifold. Convergence of Cayley SGD is theoretically analyzed. Our experiments for CNN training demonstrate that both algorithms: (a) Use less running time per iteration relative to existing approaches that enforce orthonormality of CNN parameters; and (b) Achieve faster convergence rates than the baseline SGD and ADAM algorithms without compromising the performance of the CNN. Cayley SGD and Cayley ADAM are also shown to reduce the training time for optimizing the unitary transition matrices in RNNs.

AMRL: Aggregated Memory For Reinforcement Learning

- Jacob Beck, Kamil Ciosek, Sam Devlin, Sebastian Tschiatschek, Cheng Zhang, Katja Hofmann
- abstract@open-review(Poster): In many partially observable scenarios, Reinforcement Learning (RL) agents must rely on long-term memory in order to learn an optimal policy. We demonstrate that using techniques from NLP and supervised learning fails at RL tasks due to stochasticity from the environment and from exploration. Utilizing our insights on the limitations of traditional memory methods in RL, we propose AMRL, a class of models that can learn better policies with greater sample efficiency and are resilient to noisy inputs. Specifically, our models use a standard memory module to summarize short-term context, and then aggregate all prior states from the standard model without respect to order. We show that this provides advantages both in terms of gradient decay and signal-to-noise ratio over time. Evaluating in Minecraft and maze environments that test long-term memory, we find that our model improves average return by 19% over a baseline that has the same number of parameters and by 9% over a stronger baseline that has far more parameters.

Scalable Model Compression by Entropy Penalized Reparameterization

- Deniz Oktay, Johannes Ballé, Saurabh Singh, Abhinav Shrivastava
- abstract@open-review(Poster): We describe a simple and general neural network weight compression approach, in which the network parameters (weights and biases) are represented in a “latent” space, amounting to a reparameterization. This space is equipped with a learned probability model, which is used to impose an entropy penalty on the parameter representation during training, and to compress the representation using a simple arithmetic coder after training. Classification accuracy and model compressibility is maximized jointly, with the bitrate–accuracy trade-off specified by a hyperparameter. We evaluate the method on the MNIST, CIFAR-10 and ImageNet classification benchmarks using six distinct model architectures. Our results show that state-of-the-art model compression can be achieved in a scalable and general way without requiring complex procedures such as multi-stage training.

Dynamic Time Lag Regression: Predicting What & When

- Mandar Chandorkar, Cyril Furtlehner, Bala Poduval, Enrico Camporeale, Michele Sebag
- abstract@open-review(Poster): This paper tackles a new regression problem, called Dynamic Time-Lag Regression (DTLR), where a cause signal drives an effect signal with an unknown time delay. The motivating application, pertaining to space weather modelling, aims to predict the near-Earth solar wind speed based on estimates of the Sun's coronal magnetic field. DTLR differs from mainstream regression and from sequence-to-sequence learning in two respects: firstly, no ground truth (e.g., pairs of associated sub-sequences) is available; secondly, the cause signal contains much information irrelevant to the effect signal (the solar magnetic field governs the solar wind propagation in the heliosphere, of which the Earth's magnetosphere is but a minuscule region).

A Bayesian approach is presented to tackle the specifics of the DTLR problem, with theoretical justifications based on linear stability analysis. A proof of concept on synthetic problems is presented. Finally, the empirical results on the solar wind modelling task improve on the state of the art in solar wind forecasting.

Semi-Supervised Generative Modeling for Controllable Speech Synthesis

- Raza Habib, Soroosh Mariooryad, Matt Shannon, Eric Battenberg, RJ Skerry-Ryan, Daisy Stanton, David Kao, Tom Bagby
- abstract@open-review(Poster): We present a novel generative model that combines state-of-the-art neural text- to-speech (TTS) with semi-supervised probabilistic latent variable models. By providing partial supervision to some of the latent variables, we are able to force them to take on consistent and interpretable purposes, which previously hasn’t been possible with purely unsupervised methods. We demonstrate that our model is able to reliably discover and control important but rarely labelled attributes of speech, such as affect and speaking rate, with as little as 1% (30 minutes) supervision. Even at such low supervision levels we do not observe a degradation of synthesis quality compared to a state-of-the-art baseline. We will release audio samples at https://google.github.io/tacotron/publications/semisupervised_generative_modeling_for_controllable_speech_synthesis/.

Neural Text Generation With Unlikelihood Training

- Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, Jason Weston
- abstract@open-review(Poster): Neural text generation is a key tool in natural language applications, but it is well known there are major problems at its core. In particular, standard likelihood training and decoding leads to dull and repetitive outputs. While some post-hoc fixes have been proposed, in particular top-k and nucleus sampling, they do not address the fact that the token-level probabilities predicted by the model are poor. In this paper we show that the likelihood objective itself is at fault, resulting in a model that assigns too much probability to sequences containing repeats and frequent words, unlike those from the human training distribution. We propose a new objective, unlikelihood training, which forces unlikely generations to be assigned lower probability by the model. We show that both token and sequence level unlikelihood training give less repetitive, less dull text while maintaining perplexity, giving superior generations using standard greedy or beam search. According to human evaluations, our approach with standard beam search also outperforms the currently popular decoding methods of nucleus sampling or beam blocking, thus providing a strong alternative to existing techniques.

Pure and Spurious Critical Points: a Geometric Study of Linear Networks

- Matthew Trager, Kathlén Kohn, Joan Bruna
- abstract@open-review(Poster): The critical locus of the loss function of a neural network is determined by the geometry of the functional space and by the parameterization of this space by the network's weights. We introduce a natural distinction between pure critical points, which only depend on the functional space, and spurious critical points, which arise from the parameterization. We apply this perspective to revisit and extend the literature on the loss function of linear neural networks. For this type of network, the functional space is either the set of all linear maps from input to output space, or a determinantal variety, i.e., a set of linear maps with bounded rank. We use geometric properties of determinantal varieties to derive new results on the landscape of linear networks with different loss functions and different parameterizations. Our analysis clearly illustrates that the absence of "bad" local minima in the loss landscape of linear networks is due to two distinct phenomena that apply in different settings: it is true for arbitrary smooth convex losses in the case of architectures that can express all linear maps ("filling architectures") but it holds only for the quadratic loss when the functional space is a determinantal variety ("non-filling architectures"). Without any assumption on the architecture, smooth convex losses may lead to landscapes with many bad minima.

Learning Heuristics for Quantified Boolean Formulas through Reinforcement Learning

- Gil Lederman, Markus Rabe, Sanjit Seshia, Edward A. Lee
- abstract@open-review(Poster): We demonstrate how to learn efficient heuristics for automated reasoning algorithms for quantified Boolean formulas through deep reinforcement learning. We focus on a backtracking search algorithm, which can already solve formulas of impressive size - up to hundreds of thousands of variables. The main challenge is to find a representation of these formulas that lends itself to making predictions in a scalable way. For a family of challenging problems, we learned a heuristic that solves significantly more formulas compared to the existing handwritten heuristics.

Deep neuroethology of a virtual rodent

- Josh Merel, Diego Aldarondo, Jesse Marshall, Yuval Tassa, Greg Wayne, Bence Olveczky
- abstract@open-review(Spotlight): Parallel developments in neuroscience and deep learning have led to mutually productive exchanges, pushing our understanding of real and artificial neural networks in sensory and cognitive systems. However, this interaction between fields is less developed in the study of motor control. In this work, we develop a virtual rodent as a platform for the grounded study of motor activity in artificial models of embodied control. We then use this platform to study motor activity across contexts by training a model to solve four complex tasks. Using methods familiar to neuroscientists, we describe the behavioral representations and algorithms employed by different layers of the network using a neuroethological approach to characterize motor activity relative to the rodent's behavior and goals. We find that the model uses two classes of representations which respectively encode the task-specific behavioral strategies and task-invariant behavioral kinematics. These representations are reflected in the sequential activity and population dynamics of neural subpopulations. Overall, the virtual rodent facilitates grounded collaborations between deep reinforcement learning and motor neuroscience.

Emergence of functional and structural properties of the head direction system by optimization of recurrent neural networks

- Christopher J. Cueva, Peter Y. Wang, Matthew Chin, Xue-Xin Wei
- abstract@open-review(Spotlight): Recent work suggests goal-driven training of neural networks can be used to model neural activity in the brain. While response properties of neurons in artificial neural networks bear similarities to those in the brain, the network architectures are often constrained to be different. Here we ask if a neural network can recover both neural representations and, if the architecture is unconstrained and optimized, also the anatomical properties of neural circuits. We demonstrate this in a system where the connectivity and the functional organization have been characterized, namely, the head direction circuit of the rodent and fruit fly. We trained recurrent neural networks (RNNs) to estimate head direction through integration of angular velocity. We found that the two distinct classes of neurons observed in the head direction system, the Compass neurons and the Shifter neurons, emerged naturally in artificial neural networks as a result of training. Furthermore, connectivity analysis and in-silico neurophysiology revealed structural and mechanistic similarities between artificial networks and the head direction system. Overall, our results show that optimization of RNNs in a goal-driven task can recapitulate the structure and function of biological circuits, suggesting that artificial neural networks can be used to study the brain at the level of both neural activity and anatomical organization.

Duration-of-Stay Storage Assignment under Uncertainty

- Michael Lingzhi Li, Elliott Wolf, Daniel Wintz
- abstract@open-review(Spotlight): Storage assignment, the act of choosing what goods are placed in what locations in a warehouse, is a central problem of supply chain logistics. Past literature has shown that the optimal method to assign pallets is to arrange them in increasing duration of stay in the warehouse (the Duration-of-Stay, or DoS, method), but the methodology requires perfect prior knowledge of DoS for each pallet, which is unknown and uncertain under realistic conditions. Attempts to predict DoS have largely been unfruitful due to the multi-valuedness nature (every shipment contains multiple identical pallets with different DoS) and data sparsity induced by lack of matching historical conditions. In this paper, we introduce a new framework for storage assignment that provides a solution to the DoS prediction problem through a distributional reformulation and a novel neural network, ParallelNet. Through collaboration with a world-leading cold storage company, we show that the system is able to predict DoS with a MAPE of 29%, a decrease of ~30% compared to a CNN-LSTM model, and suffers less performance decay into the future. The framework is then integrated into a first-of-its-kind Storage Assignment system, which is being deployed in warehouses across United States, with initial results showing up to 21% in labor savings. We also release the first publicly available set of warehousing records to facilitate research into this central problem.

CAQL: Continuous Action Q-Learning

- Moonkyung Ryu, Yinlam Chow, Ross Anderson, Christian Tjandraatmadja, Craig Boutilier
- abstract@open-review(Poster): Reinforcement learning (RL) with value-based methods (e.g., Q-learning) has shown success in a variety of domains such as games and recommender systems (RSs). When the action space is finite, these algorithms implicitly finds a policy by learning the optimal value function, which are often very efficient. However, one major challenge of extending Q-learning to tackle continuous-action RL problems is that obtaining optimal Bellman backup requires solving a continuous action-maximization (max-Q) problem. While it is common to restrict the parameterization of the Q-function to be concave in actions to simplify the max-Q problem, such a restriction might lead to performance degradation. Alternatively, when the Q-

function is parameterized with a generic feed-forward neural network (NN), the max-Q problem can be NP-hard. In this work, we propose the CAQL method which minimizes the Bellman residual using Q-learning with one of several plug-and-play action optimizers. In particular, leveraging the strides of optimization theories in deep NN, we show that max-Q problem can be solved optimally with mixed-integer programming (MIP)---when the Q-function has sufficient representation power, this MIP-based optimization induces better policies and is more robust than counterparts, e.g., CEM or GA, that approximate the max-Q solution. To speed up training of CAQL, we develop three techniques, namely (i) dynamic tolerance, (ii) dual filtering, and (iii) clustering. To speed up inference of CAQL, we introduce the action function that concurrently learns the optimal policy. To demonstrate the efficiency of CAQL we compare it with state-of-the-art RL algorithms on benchmark continuous control problems that have different degrees of action constraints and show that CAQL significantly outperforms policy-based methods in heavily constrained environments.

Your classifier is secretly an energy based model and you should treat it like one

- Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, Kevin Swersky
- abstract@open-review(Talk): We propose to reinterpret a standard discriminative classifier of p(y|x) as an energy based model for the joint distribution p(x, y). In this setting, the standard class probabilities can be easily computed as well as unnormalized values of p(x) and p(x|y). Within this framework, standard discriminative architectures may be used and the model can also be trained on unlabeled data. We demonstrate that energy based training of the joint distribution improves calibration, robustness, and out-of-distribution detection while also enabling our models to generate samples rivaling the quality of recent GAN approaches. We improve upon recently proposed techniques for scaling up the training of energy based models and present an approach which adds little overhead compared to standard classification training. Our approach is the first to achieve performance rivaling the state-of-the-art in both generative and discriminative learning within one hybrid model.

Adaptive Structural Fingerprints for Graph Attention Networks

- Kai Zhang, Yaokang Zhu, Jun Wang, Jie Zhang
- abstract@open-review(Poster): Graph attention network (GAT) is a promising framework to perform convolution and massage passing on graphs. Yet, how to fully exploit rich structural information in the attention mechanism remains a challenge. In the current version, GAT calculates attention scores mainly using node features and among one-hop neighbors, while increasing the attention range to higher-order neighbors can negatively affect its performance, reflecting the over-smoothing risk of GAT (or graph neural networks in general), and the ineffectiveness in exploiting graph structural details. In this paper, we propose an `"adaptive structural fingerprint" (ADSF) model to fully exploit graph topological details in graph attention network. The key idea is to contextualize each node with a weighted, learnable receptive field encoding rich and diverse local graph structures. By doing this, structural interactions between the nodes can be inferred accurately, thus significantly improving subsequent attention layer as well as the convergence of learning. Furthermore, our model provides a useful platform for different subspaces of node features and various scales of graph structures tocross-talk" with each other through the learning of multi-head attention, being particularly useful in handling complex real-world data. Empirical results demonstrate the power of our approach in exploiting rich structural information in GAT and in alleviating the intrinsic oversmoothing problem in graph neural networks.

Inductive Matrix Completion Based on Graph Neural Networks

- Muhan Zhang, Yixin Chen
- abstract@open-review(Spotlight): We propose an inductive matrix completion model without using side information. By factorizing the (rating) matrix into the product of low-dimensional latent embeddings of rows (users) and columns (items), a majority of existing matrix completion methods are transductive, since the learned embeddings cannot generalize to unseen rows/columns or to new matrices. To make matrix completion inductive, most previous works use content (side information), such as user's age or movie's genre, to make predictions. However, high-quality content is not always available, and can be hard to extract. Under the extreme setting where not any side information is available other than the matrix to complete, can we still learn an inductive matrix completion model? In this paper, we propose an Inductive Graph-based Matrix Completion (IGMC) model to address this problem. IGMC trains a graph neural network (GNN) based purely on 1-hop subgraphs around (user, item) pairs generated from the rating matrix and maps these subgraphs to their corresponding ratings. It achieves highly competitive performance with state-of-the-art transductive baselines. In addition, IGMC is inductive -- it can generalize to users/items unseen during the training (given that their interactions exist), and can even transfer to new tasks. Our transfer learning experiments show that a model trained out of the MovieLens dataset can be directly used to predict Douban movie ratings with surprisingly good performance. Our work demonstrates that: 1) it is possible to train inductive matrix completion models without using side information while achieving similar or better performances than state-of-the-art transductive methods; 2) local graph patterns around a (user, item) pair are effective predictors of the rating this user gives to the item; and 3) Long-range dependencies might not be necessary for modeling recommender systems.

ReMixMatch: Semi-Supervised Learning with Distribution Matching and Augmentation Anchoring

- David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, Colin Raffel
- abstract@open-review(Poster): We improve the recently-proposed ``MixMatch semi-supervised learning algorithm by introducing two new techniques: distribution alignment and augmentation anchoring.
- Distribution alignment encourages the marginal distribution of predictions on unlabeled data to be close to the marginal distribution of ground-truth labels.
- Augmentation anchoring} feeds multiple strongly augmented versions of an input into the model and encourages each output to be close to the prediction for a weakly-augmented version of the same input. To produce strong augmentations, we propose a variant of AutoAugment which learns the augmentation policy while the model is being trained.

Our new algorithm, dubbed ReMixMatch, is significantly more data-efficient than prior work, requiring between 5 times and 16 times less data to reach the same accuracy. For example, on CIFAR-10 with 250 labeled examples we reach 93.73% accuracy (compared to MixMatch's accuracy of 93.58% with 4000 examples) and a median accuracy of 84.92% with just four labels per class.

Conditional Learning of Fair Representations

- Han Zhao, Amanda Coston, Tameem Adel, Geoffrey J. Gordon
- abstract@open-review(Spotlight): We propose a novel algorithm for learning fair representations that can simultaneously mitigate two notions of disparity among different demographic subgroups in the classification setting. Two key components underpinning the design of our algorithm are balanced error rate and conditional alignment of representations. We show how these two components contribute to ensuring accuracy parity and equalized false-positive and false-negative rates across groups without impacting demographic parity. Furthermore, we also demonstrate both in theory and on two real-world experiments that the proposed algorithm leads to a better utility-fairness trade-off on balanced datasets compared with existing algorithms on learning fair representations for classification.

Identity Crisis: Memorization and Generalization Under Extreme Overparameterization

- Chiyuan Zhang, Samy Bengio, Moritz Hardt, Michael C. Mozer, Yoram Singer
- abstract@open-review(Poster): We study the interplay between memorization and generalization of overparameterized networks in the extreme case of a single training example and an identity-mapping task. We examine fully-connected and convolutional networks (FCN and CNN), both linear and nonlinear, initialized randomly and then trained to minimize the reconstruction error. The trained networks stereotypically take one of two forms: the

constant function (memorization) and the identity function (generalization). We formally characterize generalization in single-layer FCNs and CNNs. We show empirically that different architectures exhibit strikingly different inductive biases. For example, CNNs of up to 10 layers are able to generalize from a single example, whereas FCNs cannot learn the identity function reliably from 60k examples. Deeper CNNs often fail, but nonetheless do astonishing work to memorize the training output: because CNN biases are location invariant, the model must progressively grow an output pattern from the image boundaries via the coordination of many layers. Our work helps to quantify and visualize the sensitivity of inductive biases to architectural choices such as depth, kernel width, and number of channels.

Improved Sample Complexities for Deep Neural Networks and Robust Classification via an All-Layer Margin

- Colin Wei, Tengyu Ma
- abstract@open-review(Poster): For linear classifiers, the relationship between (normalized) output margin and generalization is captured in a clear and simple bound – a large output margin implies good generalization. Unfortunately, for deep models, this relationship is less clear: existing analyses of the output margin give complicated bounds which sometimes depend exponentially on depth. In this work, we propose to instead analyze a new notion of margin, which we call the “all-layer margin.” Our analysis reveals that the all-layer margin has a clear and direct relationship with generalization for deep models. This enables the following concrete applications of the all-layer margin: 1) by analyzing the all-layer margin, we obtain tighter generalization bounds for neural nets which depend on Jacobian and hidden layer norms and remove the exponential dependency on depth 2) our neural net results easily translate to the adversarially robust setting, giving the first direct analysis of robust test error for deep networks, and 3) we present a theoretically inspired training algorithm for increasing the all-layer margin. Our algorithm improves both clean and adversarially robust test performance over strong baselines in practice.

Cyclical Stochastic Gradient MCMC for Bayesian Deep Learning

- Ruqi Zhang, Chunyuan Li, Jianyi Zhang, Changyou Chen, Andrew Gordon Wilson
- abstract@open-review(Talk): The posteriors over neural network weights are high dimensional and multimodal. Each mode typically characterizes a meaningfully different representation of the data. We develop Cyclical Stochastic Gradient MCMC (SG-MCMC) to automatically explore such distributions. In particular, we propose a cyclical stepsize schedule, where larger steps discover new modes, and smaller steps characterize each mode. We prove non-asymptotic convergence theory of our proposed algorithm. Moreover, we provide extensive experimental results, including ImageNet, to demonstrate the effectiveness of cyclical SG-MCMC in learning complex multimodal distributions, especially for fully Bayesian inference with modern deep neural networks.

Augmenting Genetic Algorithms with Deep Neural Networks for Exploring the Chemical Space

- AkshatKumar Nigam, Pascal Friederich, Mario Krenn, Alan Aspuru-Guzik
- abstract@open-review(Poster): Challenges in natural sciences can often be phrased as optimization problems. Machine learning techniques have recently been applied to solve such problems. One example in chemistry is the design of tailor-made organic materials and molecules, which requires efficient methods to explore the chemical space. We present a genetic algorithm (GA) that is enhanced with a neural network (DNN) based discriminator model to improve the diversity of generated molecules and at the same time steer the GA. We show that our algorithm outperforms other generative models in optimization tasks. We furthermore present a way to increase interpretability of genetic algorithms, which helped us to derive design principles

Rapid Learning or Feature Reuse? Towards Understanding the Effectiveness of MAML

- Aniruddh Raghu, Maithra Raghu, Samy Bengio, Oriol Vinyals
- abstract@open-review(Poster): An important research direction in machine learning has centered around developing meta-learning algorithms to tackle few-shot learning. An especially successful algorithm has been Model Agnostic Meta-Learning (MAML), a method that consists of two optimization loops, with the outer loop finding a meta-initialization, from which the inner loop can efficiently learn new tasks. Despite MAML's popularity, a fundamental open question remains -- is the effectiveness of MAML due to the meta-initialization being primed for rapid learning (large, efficient changes in the representations) or due to feature reuse, with the meta initialization already containing high quality features? We investigate this question, via ablation studies and analysis of the latent representations, finding that feature reuse is the dominant factor. This leads to the ANIL (Almost No Inner Loop) algorithm, a simplification of MAML where we remove the inner loop for all but the (task-specific) head of the underlying neural network. ANIL matches MAML's performance on benchmark few-shot image classification and RL and offers computational improvements over MAML. We further study the precise contributions of the head and body of the network, showing that performance on the test tasks is entirely determined by the quality of the learned features, and we can remove even the head of the network (the NIL algorithm). We conclude with a discussion of the rapid learning vs feature reuse question for meta-learning algorithms more broadly.

Imitation Learning via Off-Policy Distribution Matching

- Ilya Kostrikov, Ofir Nachum, Jonathan Tompson
- abstract@open-review(Poster): When performing imitation learning from expert demonstrations, distribution matching is a popular approach, in which one alternates between estimating distribution ratios and then using these ratios as rewards in a standard reinforcement learning (RL) algorithm. Traditionally, estimation of the distribution ratio requires on-policy data, which has caused previous work to either be exorbitantly data- inefficient or alter the original objective in a manner that can drastically change its optimum. In this work, we show how the original distribution ratio estimation objective may be transformed in a principled manner to yield a completely off-policy objective. In addition to the data-efficiency that this provides, we are able to show that this objective also renders the use of a separate RL optimization unnecessary. Rather, an imitation policy may be learned directly from this objective without the use of explicit rewards. We call the resulting algorithm ValueDICE and evaluate it on a suite of popular imitation learning benchmarks, finding that it can achieve state-of-the-art sample efficiency and performance.

Reanalysis of Variance Reduced Temporal Difference Learning

- Tengyu Xu, Zhe Wang, Yi Zhou, Yingbin Liang
- abstract@open-review(Poster): Temporal difference (TD) learning is a popular algorithm for policy evaluation in reinforcement learning, but the vanilla TD can substantially suffer from the inherent optimization variance. A variance reduced TD (VRTD) algorithm was proposed by \cite{korda2015td}, which applies the variance reduction technique directly to the online TD learning with Markovian samples. In this work, we first point out the technical errors in the analysis of VRTD in \cite{korda2015td}, and then provide a mathematically solid analysis of the non-asymptotic convergence of VRTD and its variance reduction performance. We show that VRTD is guaranteed to converge to a neighborhood of the fixed-point solution of TD at a linear convergence rate. Furthermore, the variance error (for both i.i.d.\ and Markovian sampling) and the bias error (for Markovian sampling) of VRTD are significantly reduced by the batch size of variance reduction in comparison to those of vanilla TD. As a result, the overall computational complexity of VRTD to attain a given accurate solution outperforms that of TD under Markov sampling and outperforms that of TD under i.i.d.\ sampling for a sufficiently small conditional number.

Minimizing FLOPs to Learn Efficient Sparse Representations

- Biswajit Paria, Chih-Kuan Yeh, Ian E.H. Yen, Ning Xu, Pradeep Ravikumar, Barnabás Póczos
- abstract@open-review(Poster): Deep representation learning has become one of the most widely adopted approaches for visual search, recommendation, and identification. Retrieval of such representations from a large database is however computationally challenging. Approximate methods based on learning compact representations, have been widely explored for this problem, such as locality sensitive hashing, product quantization, and PCA. In this work, in contrast to learning compact representations, we propose to learn high dimensional and sparse representations that have similar representational capacity as dense embeddings while being more efficient due to sparse matrix multiplication operations which can be much faster than dense multiplication. Following the key insight that the number of operations decreases quadratically with the sparsity of embeddings provided the non-zero entries are distributed uniformly across dimensions, we propose a novel approach to learn such distributed sparse embeddings via the use of a carefully constructed regularization function that directly minimizes a continuous relaxation of the number of floating-point operations (FLOPs) incurred during retrieval. Our experiments show that our approach is competitive to the other baselines and yields a similar or better speed-vs-accuracy tradeoff on practical datasets.

Budgeted Training: Rethinking Deep Neural Network Training Under Resource Constraints

- Mengtian Li, Ersin Yumer, Deva Ramanan
- abstract@open-review(Poster): In most practical settings and theoretical analyses, one assumes that a model can be trained until convergence. However, the growing complexity of machine learning datasets and models may violate such assumptions. Indeed, current approaches for hyper-parameter tuning and neural architecture search tend to be limited by practical resource constraints. Therefore, we introduce a formal setting for studying training under the non-asymptotic, resource-constrained regime, i.e., budgeted training. We analyze the following problem: "given a dataset, algorithm, and fixed resource budget, what is the best achievable performance?" We focus on the number of optimization iterations as the representative resource. Under such a setting, we show that it is critical to adjust the learning rate schedule according to the given budget. Among budget-aware learning schedules, we find simple linear decay to be both robust and high-performing. We support our claim through extensive experiments with state-of-the-art models on ImageNet (image classification), Kinetics (video classification), MS COCO (object detection and instance segmentation), and Cityscapes (semantic segmentation). We also analyze our results and find that the key to a good schedule is budgeted convergence, a phenomenon whereby the gradient vanishes at the end of each allowed budget. We also revisit existing approaches for fast convergence and show that budget-aware learning schedules readily outperform such approaches under (the practical but under-explored) budgeted training setting.

Deep Semi-Supervised Anomaly Detection

- Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, Marius Kloft
- abstract@open-review(Poster): Deep approaches to anomaly detection have recently shown promising results over shallow methods on large and complex datasets. Typically anomaly detection is treated as an unsupervised learning problem. In practice however, one may have---in addition to a large set of unlabeled samples---access to a small pool of labeled samples, e.g. a subset verified by some domain expert as being normal or anomalous. Semi-supervised approaches to anomaly detection aim to utilize such labeled samples, but most proposed methods are limited to merely including labeled normal samples. Only a few methods take advantage of labeled anomalies, with existing deep approaches being domain-specific. In this work we present Deep SAD, an end-to-end deep methodology for general semi-supervised anomaly detection. We further introduce an information-theoretic framework for deep anomaly detection based on the idea that the entropy of the latent distribution for normal data should be lower than the entropy of the anomalous distribution, which can serve as a theoretical interpretation for our method. In extensive experiments on MNIST, Fashion-MNIST, and CIFAR-10, along with other anomaly detection benchmark datasets, we demonstrate that our method is on par or outperforms shallow, hybrid, and deep competitors, yielding appreciable performance improvements even when provided with only little labeled data.

Mirror-Generative Neural Machine Translation

- Zaixiang Zheng, Hao Zhou, Shujian Huang, Lei Li, Xin-Yu Dai, Jiajun Chen
- abstract@open-review(Talk): Training neural machine translation models (NMT) requires a large amount of parallel corpus, which is scarce for many language pairs. However, raw non-parallel corpora are often easy to obtain. Existing approaches have not exploited the full potential of non-parallel bilingual data either in training or decoding. In this paper, we propose the mirror-generative NMT (MGNMT), a single unified architecture that simultaneously integrates the source to target translation model, the target to source translation model, and two language models. Both translation models and language models share the same latent semantic space, therefore both translation directions can learn from non-parallel data more effectively. Besides, the translation models and language models can collaborate together during decoding. Our experiments show that the proposed MGNMT consistently outperforms existing approaches in a variety of scenarios and language pairs, including resource-rich and low-resource situations.

Sign Bits Are All You Need for Black-Box Attacks

- Abdullah Al-Dujaili, Una-May O'Reilly
- abstract@open-review(Poster): We present a novel black-box adversarial attack algorithm with state-of-the-art model evasion rates for query efficiency under ℓ_∞ and ℓ_2 metrics. It exploits a \textit{sign-based}, rather than magnitude-based, gradient estimation approach that shifts the gradient estimation from continuous to binary black-box optimization. It adaptively constructs queries to estimate the gradient, one query relying upon the previous, rather than re-estimating the gradient each step with random query construction. Its reliance on sign bits yields a smaller memory footprint and it requires neither hyperparameter tuning or dimensionality reduction. Further, its theoretical performance is guaranteed and it can characterize adversarial subspaces better than white-box gradient-aligned subspaces. On two public black-box attack challenges and a model robustly trained against transfer attacks, the algorithm's evasion rates surpass all submitted attacks. For a suite of published models, the algorithm is $3.8\times$ less failure-prone while spending $2.5\times$ fewer queries versus the best combination of state of art algorithms. For example, it evades a standard MNIST model using just 12 queries on average. Similar performance is observed on a standard IMAGENET model with an average of 579 queries.

Learning Hierarchical Discrete Linguistic Units from Visually-Grounded Speech

- David Harwath, *Wei-Ning Hsu*, James Glass
- abstract@open-review(Talk): In this paper, we present a method for learning discrete linguistic units by incorporating vector quantization layers into neural models of visually grounded speech. We show that our method is capable of capturing both word-level and sub-word units, depending on how it is configured. What differentiates this paper from prior work on speech unit learning is the choice of training objective. Rather than using a reconstruction-based loss, we use a discriminative, multimodal grounding objective which forces the learned units to be useful for semantic image retrieval. We evaluate the sub-word units on the ZeroSpeech 2019 challenge, achieving a 27.3% reduction in ABX error rate over the top-performing submission, while keeping the bitrate approximately the same. We also present experiments demonstrating the noise robustness of these units. Finally, we show that a model with multiple quantizers can simultaneously learn phone-like detectors at a lower layer and word-like detectors at a higher layer. We show that these detectors are highly accurate, discovering 279 words with an F1 score of greater than 0.5.

Reinforced active learning for image segmentation

- Arantxa Casanova, Pedro O. Pinheiro, Negar Rostamzadeh, Christopher J. Pal
- abstract@open-review(Poster): Learning-based approaches for semantic segmentation have two inherent challenges. First, acquiring pixel-wise labels is expensive and time-consuming. Second, realistic segmentation datasets are highly unbalanced: some categories are much more abundant than others,

biasing the performance to the most represented ones. In this paper, we are interested in focusing human labelling effort on a small subset of a larger pool of data, minimizing this effort while maximizing performance of a segmentation model on a hold-out set. We present a new active learning strategy for semantic segmentation based on deep reinforcement learning (RL). An agent learns a policy to select a subset of small informative image regions -- opposed to entire images -- to be labeled, from a pool of unlabeled data. The region selection decision is made based on predictions and uncertainties of the segmentation model being trained. Our method proposes a new modification of the deep Q-network (DQN) formulation for active learning, adapting it to the large-scale nature of semantic segmentation problems. We test the proof of concept in CamVid and provide results in the large-scale dataset Cityscapes. On Cityscapes, our deep RL region-based DQN approach requires roughly 30% less additional labeled data than our most competitive baseline to reach the same performance. Moreover, we find that our method asks for more labels of under-represented categories compared to the baselines, improving their performance and helping to mitigate class imbalance.

Gradientless Descent: High-Dimensional Zeroth-Order Optimization

- Daniel Golovin, John Karro, Greg Kochanski, Chansoo Lee, Xingyou Song, Qiuyi Zhang
- abstract@open-review(Spotlight): Zeroth-order optimization is the process of minimizing an objective $f(x)$, given oracle access to evaluations at adaptively chosen inputs x. In this paper, we present two simple yet powerful GradientLess Descent (GLD) algorithms that do not rely on an underlying gradient estimate and are numerically stable. We analyze our algorithm from a novel geometric perspective and we show that for {\it any monotone transform} of a smooth and strongly convex objective with latent dimension $k \ge n$, we present a novel analysis that shows convergence within an ϵ-ball of the optimum in $O(kQ\log(n)\log(R/\epsilon))$ evaluations, where the input dimension is n, R is the diameter of the input space and Q is the condition number. Our rates are the first of its kind to be both 1) poly-logarithmically dependent on dimensionality and 2) invariant under monotone transformations. We further leverage our geometric perspective to show that our analysis is optimal. Both monotone invariance and its ability to utilize a low latent dimensionality are key to the empirical success of our algorithms, as demonstrated on synthetic and MuJoCo benchmarks.

On the "steerability" of generative adversarial networks

- Ali Jahanian, *Lucy Chai*, Phillip Isola
- abstract@open-review(Poster): An open secret in contemporary machine learning is that many models work beautifully on standard benchmarks but fail to generalize outside the lab. This has been attributed to biased training data, which provide poor coverage over real world events. Generative models are no exception, but recent advances in generative adversarial networks (GANs) suggest otherwise -- these models can now synthesize strikingly realistic and diverse images. Is generative modeling of photos a solved problem? We show that although current GANs can fit standard datasets very well, they still fall short of being comprehensive models of the visual manifold. In particular, we study their ability to fit simple transformations such as camera movements and color changes. We find that the models reflect the biases of the datasets on which they are trained (e.g., centered objects), but that they also exhibit some capacity for generalization: by "steering" in latent space, we can shift the distribution while still creating realistic images. We hypothesize that the degree of distributional shift is related to the breadth of the training data distribution. Thus, we conduct experiments to quantify the limits of GAN transformations and introduce techniques to mitigate the problem. Code is released on our project page: https://ali-design.github.io/gan_steerability/

LEARNED STEP SIZE QUANTIZATION

- Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, Dharmendra S. Modha
- abstract@open-review(Poster): Deep networks run with low precision operations at inference time offer power and space advantages over high precision alternatives, but need to overcome the challenge of maintaining high accuracy as precision decreases. Here, we present a method for training such networks, Learned Step Size Quantization, that achieves the highest accuracy to date on the ImageNet dataset when using models, from a variety of architectures, with weights and activations quantized to 2-, 3- or 4-bits of precision, and that can train 3-bit models that reach full precision baseline accuracy. Our approach builds upon existing methods for learning weights in quantized networks by improving how the quantizer itself is configured. Specifically, we introduce a novel means to estimate and scale the task loss gradient at each weight and activation layer's quantizer step size, such that it can be learned in conjunction with other network parameters. This approach works using different levels of precision as needed for a given system and requires only a simple modification of existing training code.

Estimating counterfactual treatment outcomes over time through adversarially balanced representations

- Ioana Bica, Ahmed M Alaa, James Jordon, Mihaela van der Schaar
- abstract@open-review(Spotlight): Identifying when to give treatments to patients and how to select among multiple treatments over time are important medical problems with a few existing solutions. In this paper, we introduce the Counterfactual Recurrent Network (CRN), a novel sequence-to-sequence model that leverages the increasingly available patient observational data to estimate treatment effects over time and answer such medical questions. To handle the bias from time-varying confounders, covariates affecting the treatment assignment policy in the observational data, CRN uses domain adversarial training to build balancing representations of the patient history. At each timestep, CRN constructs a treatment invariant representation which removes the association between patient history and treatment assignments and thus can be reliably used for making counterfactual predictions. On a simulated model of tumour growth, with varying degree of time-dependent confounding, we show how our model achieves lower error in estimating counterfactuals and in choosing the correct treatment and timing of treatment than current state-of-the-art methods.

Stochastic Conditional Generative Networks with Basis Decomposition

- Ze Wang, Xiuyuan Cheng, Guillermo Sapiro, Qiang Qiu
- abstract@open-review(Poster): While generative adversarial networks (GANs) have revolutionized machine learning, a number of open questions remain to fully understand them and exploit their power. One of these questions is how to efficiently achieve proper diversity and sampling of the multi-mode data space. To address this, we introduce BasisGAN, a stochastic conditional multi-mode image generator. By exploiting the observation that a convolutional filter can be well approximated as a linear combination of a small set of basis elements, we learn a plug-and-played basis generator to stochastically generate basis elements, with just a few hundred of parameters, to fully embed stochasticity into convolutional filters. By sampling basis elements instead of filters, we dramatically reduce the cost of modeling the parameter space with no sacrifice on either image diversity or fidelity. To illustrate this proposed plug-and-play framework, we construct variants of BasisGAN based on state-of-the-art conditional image generation networks, and train the networks by simply plugging in a basis generator, without additional auxiliary components, hyperparameters, or training objectives. The experimental success is complemented with theoretical results indicating how the perturbations introduced by the proposed sampling of basis elements can propagate to the appearance of generated images.

Language GANs Falling Short

- Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, Laurent Charlin
- abstract@open-review(Poster): Traditional natural language generation (NLG) models are trained using maximum likelihood estimation (MLE) which differs from the sample generation inference procedure. During training the ground truth tokens are passed to the model, however, during inference, the model instead reads its previously generated samples - a phenomenon coined exposure bias. Exposure bias was hypothesized to be a root cause of poor sample quality and thus many generative adversarial networks (GANs) were proposed as a remedy since they have identical training and inference. However, many of the ensuing GAN variants validated sample quality improvements but ignored loss of sample diversity. This work reiterates the fallacy of quality-only metrics and clearly demonstrate that the well-established technique of reducing softmax temperature can outperform GANs on a quality-

only metric. Further, we establish a definitive quality-diversity evaluation procedure using temperature tuning over local and global sample metrics. Under this, we find that MLE models consistently outperform the proposed GAN variants over the whole quality-diversity space. Specifically, we find that 1) exposure bias appears to be less of an issue than the complications arising from non-differentiable, sequential GAN training; 2) MLE trained models provide a better quality/diversity trade-off compared to their GAN counterparts, all while being easier to train, easier to cross-validate, and less computationally expensive.

GENESIS: Generative Scene Inference and Sampling with Object-Centric Latent Representations

- Martin Engelcke, Adam R. Kosiorek, Oiwi Parker Jones, Ingmar Posner
- abstract@open-review(Poster): Generative latent-variable models are emerging as promising tools in robotics and reinforcement learning. Yet, even though tasks in these domains typically involve distinct objects, most state-of-the-art generative models do not explicitly capture the compositional nature of visual scenes. Two recent exceptions, MONet and IODINE, decompose scenes into objects in an unsupervised fashion. Their underlying generative processes, however, do not account for component interactions. Hence, neither of them allows for principled sampling of novel scenes. Here we present GENESIS, the first object-centric generative model of 3D visual scenes capable of both decomposing and generating scenes by capturing relationships between scene components. GENESIS parameterises a spatial GMM over images which is decoded from a set of object-centric latent variables that are either inferred sequentially in an amortised fashion or sampled from an autoregressive prior. We train GENESIS on several publicly available datasets and evaluate its performance on scene generation, decomposition, and semi-supervised learning.

CoPhy: Counterfactual Learning of Physical Dynamics

- Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, Christian Wolf
- abstract@open-review(Spotlight): Understanding causes and effects in mechanical systems is an essential component of reasoning in the physical world. This work poses a new problem of counterfactual learning of object mechanics from visual input. We develop the CoPhy benchmark to assess the capacity of the state-of-the-art models for causal physical reasoning in a synthetic 3D environment and propose a model for learning the physical dynamics in a counterfactual setting. Having observed a mechanical experiment that involves, for example, a falling tower of blocks, a set of bouncing balls or colliding objects, we learn to predict how its outcome is affected by an arbitrary intervention on its initial conditions, such as displacing one of the objects in the scene. The alternative future is predicted given the altered past and a latent representation of the confounders learned by the model in an end-to-end fashion with no supervision. We compare against feedforward video prediction baselines and show how observing alternative experiences allows the network to capture latent physical properties of the environment, which results in significantly more accurate predictions at the level of super human performance.

Understanding the Limitations of Variational Mutual Information Estimators

- Jiaming Song, Stefano Ermon
- abstract@open-review(Poster): Variational approaches based on neural networks are showing promise for estimating mutual information (MI) between high dimensional variables. However, they can be difficult to use in practice due to poorly understood bias/variance tradeoffs. We theoretically show that, under some conditions, estimators such as MINE exhibit variance that could grow exponentially with the true amount of underlying MI. We also empirically demonstrate that existing estimators fail to satisfy basic self-consistency properties of MI, such as data processing and additivity under independence. Based on a unified perspective of variational approaches, we develop a new estimator that focuses on variance reduction. Empirical results on standard benchmark tasks demonstrate that our proposed estimator exhibits improved bias-variance trade-offs on standard benchmark tasks.

Hamiltonian Generative Networks

- Peter Toth, Danilo J. Rezende, Andrew Jaegle, Sébastien Racanière, Aleksandar Botev, Irina Higgins
- abstract@open-review(Spotlight): The Hamiltonian formalism plays a central role in classical and quantum physics. Hamiltonians are the main tool for modelling the continuous time evolution of systems with conserved quantities, and they come equipped with many useful properties, like time reversibility and smooth interpolation in time. These properties are important for many machine learning problems - from sequence prediction to reinforcement learning and density modelling - but are not typically provided out of the box by standard tools such as recurrent neural networks. In this paper, we introduce the Hamiltonian Generative Network (HGN), the first approach capable of consistently learning Hamiltonian dynamics from high-dimensional observations (such as images) without restrictive domain assumptions. Once trained, we can use HGN to sample new trajectories, perform rollouts both forward and backward in time, and even speed up or slow down the learned dynamics. We demonstrate how a simple modification of the network architecture turns HGN into a powerful normalising flow model, called Neural Hamiltonian Flow (NHF), that uses Hamiltonian dynamics to model expressive densities. Hence, we hope that our work serves as a first practical demonstration of the value that the Hamiltonian formalism can bring to machine learning. More results and video evaluations are available at: http://tiny.cc/hgn

Feature Interaction Interpretability: A Case for Explaining Ad-Recommendation Systems via Neural Interaction Detection

- Michael Tsang, Dehua Cheng, Hanpeng Liu, Xue Feng, Eric Zhou, Yan Liu
- abstract@open-review(Poster): Recommendation is a prevalent application of machine learning that affects many users; therefore, it is important for recommender models to be accurate and interpretable. In this work, we propose a method to both interpret and augment the predictions of black-box recommender systems. In particular, we propose to interpret feature interactions from a source recommender model and explicitly encode these interactions in a target recommender model, where both source and target models are black-boxes. By not assuming the structure of the recommender system, our approach can be used in general settings. In our experiments, we focus on a prominent use of machine learning recommendation: ad-click prediction. We found that our interaction interpretations are both informative and predictive, e.g., significantly outperforming existing recommender models. What's more, the same approach to interpret interactions can provide new insights into domains even beyond recommendation, such as text and image classification.

Unsupervised Model Selection for Variational Disentangled Representation Learning

- Sunny Duan, Loic Matthey, Andre Saraiva, Nick Watters, Chris Burgess, Alexander Lerchner, Irina Higgins
- abstract@open-review(Poster): Disentangled representations have recently been shown to improve fairness, data efficiency and generalisation in simple supervised and reinforcement learning tasks. To extend the benefits of disentangled representations to more complex domains and practical applications, it is important to enable hyperparameter tuning and model selection of existing unsupervised approaches without requiring access to ground truth attribute labels, which are not available for most datasets. This paper addresses this problem by introducing a simple yet robust and reliable method for unsupervised disentangled model selection. We show that our approach performs comparably to the existing supervised alternatives across 5400 models from six state of the art unsupervised disentangled representation learning model classes. Furthermore, we show that the ranking produced by our approach correlates well with the final task performance on two different domains.

How much Position Information Do Convolutional Neural Networks Encode?

- Md Amirul Islam, *Sen Jia*, Neil D. B. Bruce

- abstract@open-review(Spotlight): In contrast to fully connected networks, Convolutional Neural Networks (CNNs) achieve efficiency by learning weights associated with local filters with a finite spatial extent. An implication of this is that a filter may know what it is looking at, but not where it is positioned in the image. Information concerning absolute position is inherently useful, and it is reasonable to assume that deep CNNs may implicitly learn to encode this information if there is a means to do so. In this paper, we test this hypothesis revealing the surprising degree of absolute position information that is encoded in commonly used neural networks. A comprehensive set of experiments show the validity of this hypothesis and shed light on how and where this information is represented while offering clues to where positional information is derived from in deep CNNs.

A Theoretical Analysis of the Number of Shots in Few-Shot Learning

- Tianshi Cao, Marc T Law, Sanja Fidler
- abstract@open-review(Poster): Few-shot classification is the task of predicting the category of an example from a set of few labeled examples. The number of labeled examples per category is called the number of shots (or shot number). Recent works tackle this task through meta-learning, where a meta-learner extracts information from observed tasks during meta-training to quickly adapt to new tasks during meta-testing. In this formulation, the number of shots exploited during meta-training has an impact on the recognition performance at meta-test time. Generally, the shot number used in meta-training should match the one used in meta-testing to obtain the best performance. We introduce a theoretical analysis of the impact of the shot number on Prototypical Networks, a state-of-the-art few-shot classification method. From our analysis, we propose a simple method that is robust to the choice of shot number used during meta-training, which is a crucial hyperparameter. The performance of our model trained for an arbitrary meta-training shot number shows great performance for different values of meta-testing shot numbers. We experimentally demonstrate our approach on different few-shot classification benchmarks.

Dynamical Distance Learning for Semi-Supervised and Unsupervised Skill Discovery

- Kristian Hartikainen, Xinyang Geng, Tuomas Haarnoja, Sergey Levine
- abstract@open-review(Poster): Reinforcement learning requires manual specification of a reward function to learn a task. While in principle this reward function only needs to specify the task goal, in practice reinforcement learning can be very time-consuming or even infeasible unless the reward function is shaped so as to provide a smooth gradient towards a successful outcome. This shaping is difficult to specify by hand, particularly when the task is learned from raw observations, such as images. In this paper, we study how we can automatically learn dynamical distances: a measure of the expected number of time steps to reach a given goal state from any other state. These dynamical distances can be used to provide well-shaped reward functions for reaching new goals, making it possible to learn complex tasks efficiently. We show that dynamical distances can be used in a semi-supervised regime, where unsupervised interaction with the environment is used to learn the dynamical distances, while a small amount of preference supervision is used to determine the task goal, without any manually engineered reward function or goal examples. We evaluate our method both on a real-world robot and in simulation. We show that our method can learn to turn a valve with a real-world 9-DoF hand, using raw image observations and just ten preference labels, without any other supervision. Videos of the learned skills can be found on the project website: https://sites.google.com/view/dynamical-distance-learning

On the Variance of the Adaptive Learning Rate and Beyond

- Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, Jiawei Han
- abstract@open-review(Poster): The learning rate warmup heuristic achieves remarkable success in stabilizing training, accelerating convergence and improving generalization for adaptive stochastic optimization algorithms like RMSprop and Adam. Pursuing the theory behind warmup, we identify a problem of the adaptive learning rate -- its variance is problematically large in the early stage, and presume warmup works as a variance reduction technique. We provide both empirical and theoretical evidence to verify our hypothesis. We further propose Rectified Adam (RAdam), a novel variant of Adam, by introducing a term to rectify the variance of the adaptive learning rate. Experimental results on image classification, language modeling, and neural machine translation verify our intuition and demonstrate the efficacy and robustness of RAdam.

Quantifying the Cost of Reliable Photo Authentication via High-Performance Learned Lossy Representations

- Pawel Korus, Nasir Memon
- abstract@open-review(Poster): Detection of photo manipulation relies on subtle statistical traces, notoriously removed by aggressive lossy compression employed online. We demonstrate that end-to-end modeling of complex photo dissemination channels allows for codec optimization with explicit provenance objectives. We design a lightweight trainable lossy image codec, that delivers competitive rate-distortion performance, on par with best hand-engineered alternatives, but has lower computational footprint on modern GPU-enabled platforms. Our results show that significant improvements in manipulation detection accuracy are possible at fractional costs in bandwidth/storage. Our codec improved the accuracy from 37% to 86% even at very low bit-rates, well below the practicality of JPEG (QF 20).

Sliced Cramer Synaptic Consolidation for Preserving Deeply Learned Representations

- Soheil Kolouri, Nicholas A. Ketz, Andrea Soltoggio, Praveen K. Pilly
- abstract@open-review(Spotlight): Deep neural networks suffer from the inability to preserve the learned data representation (i.e., catastrophic forgetting) in domains where the input data distribution is non-stationary, and it changes during training. Various selective synaptic plasticity approaches have been recently proposed to preserve network parameters, which are crucial for previously learned tasks while learning new tasks. We explore such selective synaptic plasticity approaches through a unifying lens of memory replay and show the close relationship between methods like Elastic Weight Consolidation (EWC) and Memory-Aware-Synapses (MAS). We then propose a fundamentally different class of preservation methods that aim at preserving the distribution of internal neural representations for previous tasks while learning a new one. We propose the sliced Cram\'{e}r distance as a suitable choice for such preservation and evaluate our Sliced Cramer Preservation (SCP) algorithm through extensive empirical investigations on various network architectures in both supervised and unsupervised learning settings. We show that SCP consistently utilizes the learning capacity of the network better than online-EWC and MAS methods on various incremental learning tasks.

Option Discovery using Deep Skill Chaining

- Akhil Bagaria, George Konidaris
- abstract@open-review(Poster): Autonomously discovering temporally extended actions, or skills, is a longstanding goal of hierarchical reinforcement learning. We propose a new algorithm that combines skill chaining with deep neural networks to autonomously discover skills in high-dimensional, continuous domains. The resulting algorithm, deep skill chaining, constructs skills with the property that executing one enables the agent to execute another. We demonstrate that deep skill chaining significantly outperforms both non-hierarchical agents and other state-of-the-art skill discovery techniques in challenging continuous control tasks.

HOPPITY: LEARNING GRAPH TRANSFORMATIONS TO DETECT AND FIX BUGS IN PROGRAMS

- Elizabeth Dinella, Hanjun Dai, Ziyang Li, Mayur Naik, Le Song, Ke Wang
- abstract@open-review(Spotlight): We present a learning-based approach to detect and fix a broad range of bugs in Javascript programs. We frame the problem in terms of learning a sequence of graph transformations: given a buggy program modeled by a graph structure, our model makes a sequence of

predictions including the position of bug nodes and corresponding graph edits to produce a fix. Unlike previous works that use deep neural networks, our approach targets bugs that are more complex and semantic in nature (i.e.~bugs that require adding or deleting statements to fix). We have realized our approach in a tool called HOPPITY. By training on 290,715 Javascript code change commits on Github, HOPPITY correctly detects and fixes bugs in 9,490 out of 36,361 programs in an end-to-end fashion. Given the bug location and type of the fix, HOPPITY also outperforms the baseline approach by a wide margin.

V4D: 4D Convolutional Neural Networks for Video-level Representation Learning

- Shiwen Zhang, Sheng Guo, Weilin Huang, Matthew R. Scott, Limin Wang
- abstract@open-review(Poster): Most existing 3D CNN structures for video representation learning are clip-based methods, and do not consider video-level temporal evolution of spatio-temporal features. In this paper, we propose Video-level 4D Convolutional Neural Networks, namely V4D, to model the evolution of long-range spatio-temporal representation with 4D convolutions, as well as preserving 3D spatio-temporal representations with residual connections. We further introduce the training and inference methods for the proposed V4D. Extensive experiments are conducted on three video recognition benchmarks, where V4D achieves excellent results, surpassing recent 3D CNNs by a large margin.

Learning to Represent Programs with Property Signatures

- Augustus Odena, Charles Sutton
- abstract@open-review(Poster): We introduce the notion of property signatures, a representation for programs and program specifications meant for consumption by machine learning algorithms. Given a function with input type τ_in and output type τ_out, a property is a function of type: (τ_in, τ_out) → Bool that (informally) describes some simple property of the function under consideration. For instance, if τ_in and τ_out are both lists of the same type, one property might ask ‘is the input list the same length as the output list?’. If we have a list of such properties, we can evaluate them all for our function to get a list of outputs that we will call the property signature. Crucially, we can ‘guess’ the property signature for a function given only a set of input/output pairs meant to specify that function. We discuss several potential applications of property signatures and show experimentally that they can be used to improve over a baseline synthesizer so that it emits twice as many programs in less than one-tenth of the time.

Generative Ratio Matching Networks

- Akash Srivastava, Kai Xu, Michael U. Gutmann, Charles Sutton
- abstract@open-review(Poster): Deep generative models can learn to generate realistic-looking images, but many of the most effective methods are adversarial and involve a saddlepoint optimization, which requires a careful balancing of training between a generator network and a critic network. Maximum mean discrepancy networks (MMD-nets) avoid this issue by using kernel as a fixed adversary, but unfortunately, they have not on their own been able to match the generative quality of adversarial training. In this work, we take their insight of using kernels as fixed adversaries further and present a novel method for training deep generative models that does not involve saddlepoint optimization. We call our method generative ratio matching or GRAM for short. In GRAM, the generator and the critic networks do not play a zero-sum game against each other, instead, they do so against a fixed kernel. Thus GRAM networks are not only stable to train like MMD-nets but they also match and beat the generative quality of adversarially trained generative networks.

In Search for a SAT-friendly Binarized Neural Network Architecture

- Nina Narodytska, Hongce Zhang, Aarti Gupta, Toby Walsh
- abstract@open-review(Poster): Analyzing the behavior of neural networks is one of the most pressing challenges in deep learning. Binarized Neural Networks are an important class of networks that allow equivalent representation in Boolean logic and can be analyzed formally with logic-based reasoning tools like SAT solvers. Such tools can be used to answer existential and probabilistic queries about the network, perform explanation generation, etc. However, the main bottleneck for all methods is their ability to reason about large BNNs efficiently. In this work, we analyze architectural design choices of BNNs and discuss how they affect the performance of logic-based reasoners. We propose changes to the BNN architecture and the training procedure to get a simpler network for SAT solvers without sacrificing accuracy on the primary task. Our experimental results demonstrate that our approach scales to larger deep neural networks compared to existing work for existential and probabilistic queries, leading to significant speed ups on all tested datasets.

Actor-Critic Provably Finds Nash Equilibria of Linear-Quadratic Mean-Field Games

- Zuyue Fu, Zhuoran Yang, Yongxin Chen, Zhaoran Wang
- abstract@open-review(Poster): We study discrete-time mean-field Markov games with infinite numbers of agents where each agent aims to minimize its ergodic cost. We consider the setting where the agents have identical linear state transitions and quadratic cost func- tions, while the aggregated effect of the agents is captured by the population mean of their states, namely, the mean-field state. For such a game, based on the Nash certainty equivalence principle, we provide sufficient conditions for the existence and uniqueness of its Nash equilibrium. Moreover, to find the Nash equilibrium, we propose a mean-field actor-critic algorithm with linear function approxima- tion, which does not require knowing the model of dynamics. Specifically, at each iteration of our algorithm, we use the single-agent actor-critic algorithm to approximately obtain the optimal policy of the each agent given the current mean- field state, and then update the mean-field state. In particular, we prove that our algorithm converges to the Nash equilibrium at a linear rate. To the best of our knowledge, this is the first success of applying model-free reinforcement learn- ing with function approximation to discrete-time mean-field Markov games with provable non-asymptotic global convergence guarantees.

The asymptotic spectrum of the Hessian of DNN throughout training

- Arthur Jacot, Franck Gabriel, Clement Hongler
- abstract@open-review(Poster): The dynamics of DNNs during gradient descent is described by the so-called Neural Tangent Kernel (NTK). In this article, we show that the NTK allows one to gain precise insight into the Hessian of the cost of DNNs: we obtain a full characterization of the asymptotics of the spectrum of the Hessian, at initialization and during training.

Tree-Structured Attention with Hierarchical Accumulation

- Xuan-Phi Nguyen, Shafiq Joty, Steven Hoi, Richard Socher
- abstract@open-review(Poster): Incorporating hierarchical structures like constituency trees has been shown to be effective for various natural language processing (NLP) tasks. However, it is evident that state-of-the-art (SOTA) sequence-based models like the Transformer struggle to encode such structures inherently. On the other hand, dedicated models like the Tree-LSTM, while explicitly modeling hierarchical structures, do not perform as efficiently as the Transformer. In this paper, we attempt to bridge this gap with Hierarchical Accumulation to encode parse tree structures into self-attention at constant time complexity. Our approach outperforms SOTA methods in four IWSLT translation tasks and the WMT'14 English-German task. It also yields improvements over Transformer and Tree-LSTM on three text classification tasks. We further demonstrate that using hierarchical priors can compensate for data shortage, and that our model prefers phrase-level attentions over token-level attentions.

Deep 3D Pan via local adaptive "t-shaped" convolutions with global and local adaptive dilations

- Juan Luis Gonzalez Bello, Munchurl Kim
- abstract@open-review(Poster): Recent advances in deep learning have shown promising results in many low-level vision tasks. However, solving the single-image-based view synthesis is still an open problem. In particular, the generation of new images at parallel camera views given a single input image is of great interest, as it enables 3D visualization of the 2D input scenery. We propose a novel network architecture to perform stereoscopic view synthesis at arbitrary camera positions along the X-axis, or “Deep 3D Pan”, with “t-shaped” adaptive kernels equipped with globally and locally adaptive dilations. Our proposed network architecture, the monster-net, is devised with a novel t-shaped adaptive kernel with globally and locally adaptive dilation, which can efficiently incorporate global camera shift into and handle local 3D geometries of the target image’s pixels for the synthesis of naturally looking 3D panned views when a 2-D input image is given. Extensive experiments were performed on the KITTI, CityScapes, and our VICLAB_STEREO indoors dataset to prove the efficacy of our method. Our monster-net significantly outperforms the state-of-the-art method (SOTA) by a large margin in all metrics of RMSE, PSNR, and SSIM. Our proposed monster-net is capable of reconstructing more reliable image structures in synthesized images with coherent geometry. Moreover, the disparity information that can be extracted from the “t-shaped” kernel is much more reliable than that of the SOTA for the unsupervised monocular depth estimation task, confirming the effectiveness of our method.

Low-Resource Knowledge-Grounded Dialogue Generation

- Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, Rui Yan
- abstract@open-review(Poster): Responding with knowledge has been recognized as an important capability for an intelligent conversational agent. Yet knowledge-grounded dialogues, as training data for learning such a response generation model, are difficult to obtain. Motivated by the challenge in practice, we consider knowledge-grounded dialogue generation under a natural assumption that only limited training examples are available. In such a low-resource setting, we devise a disentangled response decoder in order to isolate parameters that depend on knowledge-grounded dialogues from the entire generation model. By this means, the major part of the model can be learned from a large number of ungrounded dialogues and unstructured documents, while the remaining small parameters can be well fitted using the limited training examples. Evaluation results on two benchmarks indicate that with only $1/8$ training data, our model can achieve the state-of-the-art performance and generalize well on out-of-domain knowledge.

A Target-Agnostic Attack on Deep Models: Exploiting Security Vulnerabilities of Transfer Learning

- Shahbaz Rezaei, Xin Liu
- abstract@open-review(Poster): Due to insufficient training data and the high computational cost to train a deep neural network from scratch, transfer learning has been extensively used in many deep-neural-network-based applications. A commonly used transfer learning approach involves taking a part of a pre-trained model, adding a few layers at the end, and re-training the new layers with a small dataset. This approach, while efficient and widely used, imposes a security vulnerability because the pre-trained model used in transfer learning is usually publicly available, including to potential attackers. In this paper, we show that without any additional knowledge other than the pre-trained model, an attacker can launch an effective and efficient brute force attack that can craft instances of input to trigger each target class with high confidence. We assume that the attacker has no access to any target-specific information, including samples from target classes, re-trained model, and probabilities assigned by Softmax to each class, and thus making the attack target-agnostic. These assumptions render all previous attack models inapplicable, to the best of our knowledge. To evaluate the proposed attack, we perform a set of experiments on face recognition and speech recognition tasks and show the effectiveness of the attack. Our work reveals a fundamental security weakness of the Softmax layer when used in transfer learning settings.

On Computation and Generalization of Generative Adversarial Imitation Learning

- Minshuo Chen, Yizhou Wang, Tianyi Liu, Zhuoran Yang, Xingguo Li, Zhaoran Wang, Tuo Zhao
- abstract@open-review(Poster): Generative Adversarial Imitation Learning (GAIL) is a powerful and practical approach for learning sequential decision-making policies. Different from Reinforcement Learning (RL), GAIL takes advantage of demonstration data by experts (e.g., human), and learns both the policy and reward function of the unknown environment. Despite the significant empirical progresses, the theory behind GAIL is still largely unknown. The major difficulty comes from the underlying temporal dependency of the demonstration data and the minimax computational formulation of GAIL without convex-concave structure. To bridge such a gap between theory and practice, this paper investigates the theoretical properties of GAIL. Specifically, we show: (1) For GAIL with general reward parameterization, the generalization can be guaranteed as long as the class of the reward functions is properly controlled; (2) For GAIL, where the reward is parameterized as a reproducing kernel function, GAIL can be efficiently solved by stochastic first order optimization algorithms, which attain sublinear convergence to a stationary solution. To the best of our knowledge, these are the first results on statistical and computational guarantees of imitation learning with reward/policy function ap- proximation. Numerical experiments are provided to support our analysis.

FEW-SHOT LEARNING ON GRAPHS VIA SUPER-CLASSES BASED ON GRAPH SPECTRAL MEASURES

- Jatin Chauhan, Deepak Nathani, Manohar Kaul
- abstract@open-review(Poster): We propose to study the problem of few-shot graph classification in graph neural networks (GNNs) to recognize unseen classes, given limited labeled graph examples. Despite several interesting GNN variants being proposed recently for node and graph classification tasks, when faced with scarce labeled examples in the few-shot setting, these GNNs exhibit significant loss in classification performance. Here, we present an approach where a probability measure is assigned to each graph based on the spectrum of the graph’s normalized Laplacian. This enables us to accordingly cluster the graph base-labels associated with each graph into super-classes, where the L^p Wasserstein distance serves as our underlying distance metric. Subsequently, a super-graph constructed based on the super-classes is then fed to our proposed GNN framework which exploits the latent inter-class relationships made explicit by the super-graph to achieve better class label separation among the graphs. We conduct exhaustive empirical evaluations of our proposed method and show that it outperforms both the adaptation of state-of-the-art graph classification methods to few-shot scenario and our naive baseline GNNs. Additionally, we also extend and study the behavior of our method to semi-supervised and active learning scenarios.

Influence-Based Multi-Agent Exploration

- Tonghan Wang, *Jianhao Wang*, Yi Wu, Chongjie Zhang
- abstract@open-review(Spotlight): Intrinsically motivated reinforcement learning aims to address the exploration challenge for sparse-reward tasks. However, the study of exploration methods in transition-dependent multi-agent settings is largely absent from the literature. We aim to take a step towards solving this problem. We present two exploration methods: exploration via information-theoretic influence (EITI) and exploration via decision-theoretic influence (EDTI), by exploiting the role of interaction in coordinated behaviors of agents. EITI uses mutual information to capture the interdependence between the transition dynamics of agents. EDTI uses a novel intrinsic reward, called Value of Interaction (VoI), to characterize and quantify the influence of one agent's behavior on expected returns of other agents. By optimizing EITI or EDTI objective as a regularizer, agents are encouraged to coordinate their exploration and learn policies to optimize the team performance. We show how to optimize these regularizers so that they can be easily integrated with policy gradient reinforcement learning. The resulting update rule draws a connection between coordinated exploration and intrinsic reward distribution. Finally, we empirically demonstrate the significant strength of our methods in a variety of multi-agent scenarios.

Multiplicative Interactions and Where to Find Them

- Siddhant M. Jayakumar, Wojciech M. Czarnecki, Jacob Menick, Jonathan Schwarz, Jack Rae, Simon Osindero, Yee Whye Teh, Tim Harley, Razvan Pascanu
- abstract@open-review(Poster): We explore the role of multiplicative interaction as a unifying framework to describe a range of classical and modern neural network architectural motifs, such as gating, attention layers, hypernetworks, and dynamic convolutions amongst others. Multiplicative interaction layers as primitive operations have a long-established presence in the literature, though this often not emphasized and thus under-appreciated. We begin by showing that such layers strictly enrich the representable function classes of neural networks. We conjecture that multiplicative interactions offer a particularly powerful inductive bias when fusing multiple streams of information or when conditional computation is required. We therefore argue that they should be considered in many situation where multiple compute or information paths need to be combined, in place of the simple and oft-used concatenation operation. Finally, we back up our claims and demonstrate the potential of multiplicative interactions by applying them in large-scale complex RL and sequence modelling tasks, where their use allows us to deliver state-of-the-art results, and thereby provides new evidence in support of multiplicative interactions playing a more prominent role when designing new neural network architectures.

Continual Learning with Bayesian Neural Networks for Non-Stationary Data

- Richard Kurle, Botond Cseke, Alexej Klushyn, Patrick van der Smagt, Stephan Günnemann
- abstract@open-review(Poster): This work addresses continual learning for non-stationary data, using Bayesian neural networks and memory-based online variational Bayes. We represent the posterior approximation of the network weights by a diagonal Gaussian distribution and a complementary memory of raw data. This raw data corresponds to likelihood terms that cannot be well approximated by the Gaussian. We introduce a novel method for sequentially updating both components of the posterior approximation. Furthermore, we propose Bayesian forgetting and a Gaussian diffusion process for adapting to non-stationary data. The experimental results show that our update method improves on existing approaches for streaming data. Additionally, the adaptation methods lead to better predictive performance for non-stationary data.

SAdam: A Variant of Adam for Strongly Convex Functions

- Guanghui Wang, Shiyin Lu, Quan Cheng, Wei-wei Tu, Lijun Zhang
- abstract@open-review(Poster): The Adam algorithm has become extremely popular for large-scale machine learning. Under convexity condition, it has been proved to enjoy a data-dependent $O(\sqrt{T})$ regret bound where T is the time horizon. However, whether strong convexity can be utilized to further improve the performance remains an open problem. In this paper, we give an affirmative answer by developing a variant of Adam (referred to as SAdam) which achieves a data-dependent $O(\log T)$ regret bound for strongly convex functions. The essential idea is to maintain a faster decaying yet under controlled step size for exploiting strong convexity. In addition, under a special configuration of hyperparameters, our SAdam reduces to SC-RMSprop, a recently proposed variant of RMSprop for strongly convex functions, for which we provide the first data-dependent logarithmic regret bound. Empirical results on optimizing strongly convex functions and training deep networks demonstrate the effectiveness of our method.

Generalization bounds for deep convolutional neural networks

- Philip M. Long, Hanie Sedghi
- abstract@open-review(Poster): We prove bounds on the generalization error of convolutional networks. The bounds are in terms of the training loss, the number of parameters, the Lipschitz constant of the loss and the distance from the weights to the initial weights. They are independent of the number of pixels in the input, and the height and width of hidden feature maps. We present experiments using CIFAR-10 with varying hyperparameters of a deep convolutional network, comparing our bounds with practical generalization gaps.

A Fair Comparison of Graph Neural Networks for Graph Classification

- Federico Errica, Marco Podda, Davide Bacciu, Alessio Micheli
- abstract@open-review(Poster): Experimental reproducibility and replicability are critical topics in machine learning. Authors have often raised concerns about their lack in scientific publications to improve the quality of the field. Recently, the graph representation learning field has attracted the attention of a wide research community, which resulted in a large stream of works. As such, several Graph Neural Network models have been developed to effectively tackle graph classification. However, experimental procedures often lack rigorousness and are hardly reproducible. Motivated by this, we provide an overview of common practices that should be avoided to fairly compare with the state of the art. To counter this troubling trend, we ran more than 47000 experiments in a controlled and uniform framework to re-evaluate five popular models across nine common benchmarks. Moreover, by comparing GNNs with structure-agnostic baselines we provide convincing evidence that, on some datasets, structural information has not been exploited yet. We believe that this work can contribute to the development of the graph learning field, by providing a much needed grounding for rigorous evaluations of graph classification models.

Finding and Visualizing Weaknesses of Deep Reinforcement Learning Agents

- Christian Rupprecht, Cyril Ibrahim, Christopher J. Pal
- abstract@open-review(Poster): As deep reinforcement learning driven by visual perception becomes more widely used there is a growing need to better understand and probe the learned agents. Understanding the decision making process and its relationship to visual inputs can be very valuable to identify problems in learned behavior. However, this topic has been relatively under-explored in the research community. In this work we present a method for synthesizing visual inputs of interest for a trained agent. Such inputs or states could be situations in which specific actions are necessary. Further, critical states in which a very high or a very low reward can be achieved are often interesting to understand the situational awareness of the system as they can correspond to risky states. To this end, we learn a generative model over the state space of the environment and use its latent space to optimize a target function for the state of interest. In our experiments we show that this method can generate insights for a variety of environments and reinforcement learning methods. We explore results in the standard Atari benchmark games as well as in an autonomous driving simulator. Based on the efficiency with which we have been able to identify behavioural weaknesses with this technique, we believe this general approach could serve as an important tool for AI safety applications.

Computation Reallocation for Object Detection

- Feng Liang, Chen Lin, Ronghao Guo, Ming Sun, Wei Wu, Junjie Yan, Wanli Ouyang
- abstract@open-review(Poster): The allocation of computation resources in the backbone is a crucial issue in object detection. However, classification allocation pattern is usually adopted directly to object detector, which is proved to be sub-optimal. In order to reallocate the engaged computation resources in a more efficient way, we present CR-NAS (Computation Reallocation Neural Architecture Search) that can learn computation reallocation strategies across different feature resolution and spatial position diectly on the target detection dataset. A two-level reallocation space is proposed for both stage and spatial reallocation. A novel hierarchical search procedure is adopted to cope with the complex search space. We apply CR-NAS to multiple backbones and achieve consistent improvements. Our CR-ResNet50 and CR-MobileNetV2 outperforms the baseline by 1.9% and 1.7% COCO AP respectively without any additional computation budget. The models discovered by CR-NAS can be equiped to other powerful detection neck/head and be easily transferred to other dataset, e.g. PASCAL VOC, and other vision tasks, e.g. instance segmentation. Our CR-NAS can be used as a plugin to improve the performance of various networks, which is demanding.

Meta-Learning without Memorization

- Mingzhang Yin, George Tucker, Mingyuan Zhou, Sergey Levine, Chelsea Finn
- abstract@open-review(Spotlight): The ability to learn new concepts with small amounts of data is a critical aspect of intelligence that has proven challenging for deep learning methods. Meta-learning has emerged as a promising technique for leveraging data from previous tasks to enable efficient learning of new tasks. However, most meta-learning algorithms implicitly require that the meta-training tasks be mutually-exclusive, such that no single model can solve all of the tasks at once. For example, when creating tasks for few-shot image classification, prior work uses a per-task random assignment of image classes to N-way classification labels. If this is not done, the meta-learner can ignore the task training data and learn a single model that performs all of the meta-training tasks zero-shot, but does not adapt effectively to new image classes. This requirement means that the user must take great care in designing the tasks, for example by shuffling labels or removing task identifying information from the inputs. In some domains, this makes meta-learning entirely inapplicable. In this paper, we address this challenge by designing a meta-regularization objective using information theory that places precedence on data-driven adaptation. This causes the meta-learner to decide what must be learned from the task training data and what should be inferred from the task testing input. By doing so, our algorithm can successfully use data from non-mutually-exclusive tasks to efficiently adapt to novel tasks. We demonstrate its applicability to both contextual and gradient-based meta-learning algorithms, and apply it in practical settings where applying standard meta-learning has been difficult. Our approach substantially outperforms standard meta-learning algorithms in these settings.

From Variational to Deterministic Autoencoders

- Partha Ghosh, Mehdi S. M. Sajjadi, Antonio Vergari, Michael Black, Bernhard Scholkopf
- abstract@open-review(Poster): Variational Autoencoders (VAEs) provide a theoretically-backed and popular framework for deep generative models. However, learning a VAE from data poses still unanswered theoretical questions and considerable practical challenges. In this work, we propose an alternative framework for generative modeling that is simpler, easier to train, and deterministic, yet has many of the advantages of the VAE. We observe that sampling a stochastic encoder in a Gaussian VAE can be interpreted as simply injecting noise into the input of a deterministic decoder. We investigate how substituting this kind of stochasticity, with other explicit and implicit regularization schemes, can lead to an equally smooth and meaningful latent space without having to force it to conform to an arbitrarily chosen prior. To retrieve a generative mechanism to sample new data points, we introduce an ex-post density estimation step that can be readily applied to the proposed framework as well as existing VAEs, improving their sample quality. We show, in a rigorous empirical study, that the proposed regularized deterministic autoencoders are able to generate samples that are comparable to, or better than, those of VAEs and more powerful alternatives when applied to images as well as to structured data such as molecules.

Adversarially Robust Representations with Smooth Encoders

- Taylan Cemgil, Sumedh Ghaisas, Krishnamurthy (Dj) Dvijotham, Pushmeet Kohli
- abstract@open-review(Poster): This paper studies the undesired phenomena of over-sensitivity of representations learned by deep networks to semantically-irrelevant changes in data. We identify a cause for this shortcoming in the classical Variational Auto-encoder (VAE) objective, the evidence lower bound (ELBO). We show that the ELBO fails to control the behaviour of the encoder out of the support of the empirical data distribution and this behaviour of the VAE can lead to extreme errors in the learned representation. This is a key hurdle in the effective use of representations for data-efficient learning and transfer. To address this problem, we propose to augment the data with specifications that enforce insensitivity of the representation with respect to families of transformations. To incorporate these specifications, we propose a regularization method that is based on a selection mechanism that creates a fictive data point by explicitly perturbing an observed true data point. For certain choices of parameters, our formulation naturally leads to the minimization of the entropy regularized Wasserstein distance between representations. We illustrate our approach on standard datasets and experimentally show that significant improvements in the downstream adversarial accuracy can be achieved by learning robust representations completely in an unsupervised manner, without a reference to a particular downstream task and without a costly supervised adversarial training procedure.

AssembleNet: Searching for Multi-Stream Neural Connectivity in Video Architectures

- Michael S. Ryoo, AJ Piergiovanni, Mingxing Tan, Anelia Angelova
- abstract@open-review(Poster): Learning to represent videos is a very challenging task both algorithmically and computationally. Standard video CNN architectures have been designed by directly extending architectures devised for image understanding to include the time dimension, using modules such as 3D convolutions, or by using two-stream design to capture both appearance and motion in videos. We interpret a video CNN as a collection of multi-stream convolutional blocks connected to each other, and propose the approach of automatically finding neural architectures with better connectivity and spatio-temporal interactions for video understanding. This is done by evolving a population of overly-connected architectures guided by connection weight learning. Architectures combining representations that abstract different input types (i.e., RGB and optical flow) at multiple temporal resolutions are searched for, allowing different types or sources of information to interact with each other. Our method, referred to as AssembleNet, outperforms prior approaches on public video datasets, in some cases by a great margin. We obtain 58.6% mAP on Charades and 34.27% accuracy on Moments-in-Time.

ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning

- Weihao Yu, Zihang Jiang, Yanfei Dong, Jiashi Feng
- abstract@open-review(Poster): Recent powerful pre-trained language models have achieved remarkable performance on most of the popular datasets for reading comprehension. It is time to introduce more challenging datasets to push the development of this field towards more comprehensive reasoning of text. In this paper, we introduce a new Reading Comprehension dataset requiring logical reasoning (ReClor) extracted from standardized graduate admission examinations. As earlier studies suggest, human-annotated datasets usually contain biases, which are often exploited by models to achieve high accuracy without truly understanding the text. In order to comprehensively evaluate the logical reasoning ability of models on ReClor, we propose to identify biased data points and separate them into EASY set while the rest as HARD set. Empirical results show that state-of-the-art models have an outstanding ability to capture biases contained in the dataset with high accuracy on EASY set. However, they struggle on HARD set with poor performance near that of random guess, indicating more research is needed to essentially enhance the logical reasoning ability of current models.

Finite Depth and Width Corrections to the Neural Tangent Kernel

- Boris Hanin, Mihai Nica
- abstract@open-review(Spotlight): We prove the precise scaling, at finite depth and width, for the mean and variance of the neural tangent kernel (NTK) in a randomly initialized ReLU network. The standard deviation is exponential in the ratio of network depth to width. Thus, even in the limit of infinite overparameterization, the NTK is not deterministic if depth and width simultaneously tend to infinity. Moreover, we prove that for such deep and wide networks, the NTK has a non-trivial evolution during training by showing that the mean of its first SGD update is also exponential in the ratio of network depth to width. This is sharp contrast to the regime where depth is fixed and network width is very large. Our results suggest that, unlike relatively shallow and wide networks, deep and wide ReLU networks are capable of learning data-dependent features even in the so-called lazy training regime.

Order Learning and Its Application to Age Estimation

- Kyungsun Lim, Nyeong-Ho Shin, Young-Yoon Lee, Chang-Su Kim

- abstract@open-review(Poster): We propose order learning to determine the order graph of classes, representing ranks or priorities, and classify an object instance into one of the classes. To this end, we design a pairwise comparator to categorize the relationship between two instances into one of three cases: one instance is `greater than,'similar to,' or `smaller than' the other. Then, by comparing an input instance with reference instances and maximizing the consistency among the comparison results, the class of the input can be estimated reliably. We apply order learning to develop a facial age estimator, which provides the state-of-the-art performance. Moreover, the performance is further improved when the order graph is divided into disjoint chains using gender and ethnic group information or even in an unsupervised manner.

Efficient and Information-Preserving Future Frame Prediction and Beyond

- Wei Yu, Yichao Lu, Steve Easterbrook, Sanja Fidler
- abstract@open-review(Poster): Applying resolution-preserving blocks is a common practice to maximize information preservation in video prediction, yet their high memory consumption greatly limits their application scenarios. We propose CrevNet, a Conditionally Reversible Network that uses reversible architectures to build a bijective two-way autoencoder and its complementary recurrent predictor. Our model enjoys the theoretically guaranteed property of no information loss during the feature extraction, much lower memory consumption and computational efficiency. The lightweight nature of our model enables us to incorporate 3D convolutions without concern of memory bottleneck, enhancing the model's ability to capture both short-term and long-term temporal dependencies. Our proposed approach achieves state-of-the-art results on Moving MNIST, Traffic4cast and KITTI datasets. We further demonstrate the transferability of our self-supervised learning method by exploiting its learnt features for object detection on KITTI. Our competitive results indicate the potential of using CrevNet as a generative pre-training strategy to guide downstream tasks.

NAS evaluation is frustratingly hard

- Antoine Yang, Pedro M. Esperança, Fabio M. Carlucci
- abstract@open-review(Poster): Neural Architecture Search (NAS) is an exciting new field which promises to be as much as a game-changer as Convolutional Neural Networks were in 2012. Despite many great works leading to substantial improvements on a variety of tasks, comparison between different methods is still very much an open issue. While most algorithms are tested on the same datasets, there is no shared experimental protocol followed by all. As such, and due to the under-use of ablation studies, there is a lack of clarity regarding why certain methods are more effective than others. Our first contribution is a benchmark of 8 NAS methods on 5 datasets. To overcome the hurdle of comparing methods with different search spaces, we propose using a method's relative improvement over the randomly sampled average architecture, which effectively removes advantages arising from expertly engineered search spaces or training protocols. Surprisingly, we find that many NAS techniques struggle to significantly beat the average architecture baseline. We perform further experiments with the commonly used DARTS search space in order to understand the contribution of each component in the NAS pipeline. These experiments highlight that: (i) the use of tricks in the evaluation protocol has a predominant impact on the reported performance of architectures; (ii) the cell-based search space has a very narrow accuracy range, such that the seed has a considerable impact on architecture rankings; (iii) the hand-designed macrostructure (cells) is more important than the searched micro-structure (operations); and (iv) the depth-gap is a real phenomenon, evidenced by the change in rankings between 8 and 20 cell architectures. To conclude, we suggest best practices, that we hope will prove useful for the community and help mitigate current NAS pitfalls, e.g. difficulties in reproducibility and comparison of search methods. The code used is available at https://github.com/antoyang/NAS-Benchmark.

CLN2INV: Learning Loop Invariants with Continuous Logic Networks

- Gabriel Ryan, Justin Wong, Jianan Yao, Ronghui Gu, Suman Jana
- abstract@open-review(Poster): Program verification offers a framework for ensuring program correctness and therefore systematically eliminating different classes of bugs. Inferring loop invariants is one of the main challenges behind automated verification of real-world programs which often contain many loops. In this paper, we present the Continuous Logic Network (CLN), a novel neural architecture for automatically learning loop invariants directly from program execution traces. Unlike existing neural networks, CLNs can learn precise and explicit representations of formulas in Satisfiability Modulo Theories (SMT) for loop invariants from program execution traces. We develop a new sound and complete semantic mapping for assigning SMT formulas to continuous truth values that allows CLNs to be trained efficiently. We use CLNs to implement a new inference system for loop invariants, CLN2INV, that significantly outperforms existing approaches on the popular Code2Inv dataset. CLN2INV is the first tool to solve all 124 theoretically solvable problems in the Code2Inv dataset. Moreover, CLN2INV takes only 1.1 second on average for each problem, which is 40 times faster than existing approaches. We further demonstrate that CLN2INV can even learn 12 significantly more complex loop invariants than the ones required for the Code2Inv dataset.

Scalable Neural Methods for Reasoning With a Symbolic Knowledge Base

- William W. Cohen, Haitian Sun, R. Alex Hofer, Matthew Siegler
- abstract@open-review(Poster): We describe a novel way of representing a symbolic knowledge base (KB) called a sparse-matrix reified KB. This representation enables neural modules that are fully differentiable, faithful to the original semantics of the KB, expressive enough to model multi-hop inferences, and scalable enough to use with realistically large KBs. The sparse-matrix reified KB can be distributed across multiple GPUs, can scale to tens of millions of entities and facts, and is orders of magnitude faster than naive sparse-matrix implementations. The reified KB enables very simple end-to-end architectures to obtain competitive performance on several benchmarks representing two families of tasks: KB completion, and learning semantic parsers from denotations.

Ridge Regression: Structure, Cross-Validation, and Sketching

- Sifan Liu, Edgar Dobriban
- abstract@open-review(Spotlight): We study the following three fundamental problems about ridge regression: (1) what is the structure of the estimator? (2) how to correctly use cross-validation to choose the regularization parameter? and (3) how to accelerate computation without losing too much accuracy? We consider the three problems in a unified large-data linear model. We give a precise representation of ridge regression as a covariance matrix-dependent linear combination of the true parameter and the noise. We study the bias of K-fold cross-validation for choosing the regularization parameter, and propose a simple bias-correction. We analyze the accuracy of primal and dual sketching for ridge regression, showing they are surprisingly accurate. Our results are illustrated by simulations and by analyzing empirical data.

A Constructive Prediction of the Generalization Error Across Scales

- Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, Nir Shavit
- abstract@open-review(Poster): The dependency of the generalization error of neural networks on model and dataset size is of critical importance both in practice and for understanding the theory of neural networks. Nevertheless, the functional form of this dependency remains elusive. In this work, we present a functional form which approximates well the generalization error in practice. Capitalizing on the successful concept of model scaling (e.g., width, depth), we are able to simultaneously construct such a form and specify the exact models which can attain it across model/data scales. Our construction follows insights obtained from observations conducted over a range of model/data scales, in various model types and datasets, in vision and language tasks. We show that the form both fits the observations well across scales, and provides accurate predictions from small- to large-scale models and data.

An Inductive Bias for Distances: Neural Nets that Respect the Triangle Inequality

- Silviu Pitis, Harris Chan, Kiarash Jamali, Jimmy Ba
- abstract@open-review(Poster): Distances are pervasive in machine learning. They serve as similarity measures, loss functions, and learning targets; it is said that a good distance measure solves a task. When defining distances, the triangle inequality has proven to be a useful constraint, both theoretically---to prove convergence and optimality guarantees---and empirically---as an inductive bias. Deep metric learning architectures that respect the triangle inequality rely, almost exclusively, on Euclidean distance in the latent space. Though effective, this fails to model two broad classes of subadditive distances, common in graphs and reinforcement learning: asymmetric metrics, and metrics that cannot be embedded into Euclidean space. To address these problems, we introduce novel architectures that are guaranteed to satisfy the triangle inequality. We prove our architectures universally approximate norm-induced metrics on $\mathbb{R}^n$, and present a similar result for modified Input Convex Neural Networks. We show that our architectures outperform existing metric approaches when modeling graph distances and have a better inductive bias than non-metric approaches when training data is limited in the multi-goal reinforcement learning setting.

Mogrifier LSTM

- Gábor Melis, Tomáš Kočiský, Phil Blunsom
- abstract@open-review(Talk): Many advances in Natural Language Processing have been based upon more expressive models for how inputs interact with the context in which they occur. Recurrent networks, which have enjoyed a modicum of success, still lack the generalization and systematicity ultimately required for modelling language. In this work, we propose an extension to the venerable Long Short-Term Memory in the form of mutual gating of the current input and the previous output. This mechanism affords the modelling of a richer space of interactions between inputs and their context. Equivalently, our model can be viewed as making the transition function given by the LSTM context-dependent. Experiments demonstrate markedly improved generalization on language modelling in the range of 3–4 perplexity points on Penn Treebank and Wikitext-2, and 0.01–0.05 bpc on four character-based datasets. We establish a new state of the art on all datasets with the exception of Enwik8, where we close a large gap between the LSTM and Transformer models.

Physics-as-Inverse-Graphics: Unsupervised Physical Parameter Estimation from Video

- Miguel Jaques, Michael Burke, Timothy Hospedales
- abstract@open-review(Poster): We propose a model that is able to perform physical parameter estimation of systems from video, where the differential equations governing the scene dynamics are known, but labeled states or objects are not available. Existing physical scene understanding methods require either object state supervision, or do not integrate with differentiable physics to learn interpretable system parameters and states. We address this problem through a \textit{physics-as-inverse-graphics} approach that brings together vision-as-inverse-graphics and differentiable physics engines, where objects and explicit state and velocity representations are discovered by the model. This framework allows us to perform long term extrapolative video prediction, as well as vision-based model-predictive control. Our approach significantly outperforms related unsupervised methods in long-term future frame prediction of systems with interacting objects (such as ball-spring or 3-body gravitational systems), due to its ability to build dynamics into the model as an inductive bias. We further show the value of this tight vision-physics integration by demonstrating data-efficient learning of vision-actuated model-based control for a pendulum system. We also show that the controller's interpretability provides unique capabilities in goal-driven control and physical reasoning for zero-data adaptation.

Counterfactuals uncover the modular structure of deep generative models

- Michel Besserve, Arash Mehrjou, Rémy Sun, Bernhard Schölkopf
- abstract@open-review(Poster): Deep generative models can emulate the perceptual properties of complex image datasets, providing a latent representation of the data. However, manipulating such representation to perform meaningful and controllable transformations in the data space remains challenging without some form of supervision. While previous work has focused on exploiting statistical independence to \textit{disentangle} latent factors, we argue that such requirement can be advantageously relaxed and propose instead a non-statistical framework that relies on identifying a modular organization of the network, based on counterfactual manipulations. Our experiments support that modularity between groups of channels is achieved to a certain degree on a variety of generative models. This allowed the design of targeted interventions on complex image datasets, opening the way to applications such as computationally efficient style transfer and the automated assessment of robustness to contextual changes in pattern recognition systems.

Gap-Aware Mitigation of Gradient Staleness

- Saar Barkai, Ido Hakimi, Assaf Schuster
- abstract@open-review(Poster): Cloud computing is becoming increasingly popular as a platform for distributed training of deep neural networks. Synchronous stochastic gradient descent (SSGD) suffers from substantial slowdowns due to stragglers if the environment is non-dedicated, as is common in cloud computing. Asynchronous SGD (ASGD) methods are immune to these slowdowns but are scarcely used due to gradient staleness, which encumbers the convergence process. Recent techniques have had limited success mitigating the gradient staleness when scaling up to many workers (computing nodes). In this paper we define the Gap as a measure of gradient staleness and propose Gap-Aware (GA), a novel asynchronous-distributed method that penalizes stale gradients linearly to the Gap and performs well even when scaling to large numbers of workers. Our evaluation on the CIFAR, ImageNet, and WikiText-103 datasets shows that GA outperforms the currently acceptable gradient penalization method, in final test accuracy. We also provide convergence rate proof for GA. Despite prior beliefs, we show that if GA is applied, momentum becomes beneficial in asynchronous environments, even when the number of workers scales up.

Ensemble Distribution Distillation

- Andrey Malinin, Bruno Mlodozeniec, Mark Gales
- abstract@open-review(Poster): Ensembles of models often yield improvements in system performance. These ensemble approaches have also been empirically shown to yield robust measures of uncertainty, and are capable of distinguishing between different forms of uncertainty. However, ensembles come at a computational and memory cost which may be prohibitive for many applications. There has been significant work done on the distillation of an ensemble into a single model. Such approaches decrease computational cost and allow a single model to achieve an accuracy comparable to that of an ensemble. However, information about the diversity of the ensemble, which can yield estimates of different forms of uncertainty, is lost. This work considers the novel task of Ensemble Distribution Distillation (EnD^2) - distilling the distribution of the predictions from an ensemble, rather than just the average prediction, into a single model. EnD^2 enables a single model to retain both the improved classification performance of ensemble distillation as well as information about the diversity of the ensemble, which is useful for uncertainty estimation. A solution for EnD^2 based on Prior Networks, a class of models which allow a single neural network to explicitly model a distribution over output distributions, is proposed in this work. The properties of EnD^2 are investigated on both an artificial dataset, and on the CIFAR-10, CIFAR-100 and TinyImageNet datasets, where it is shown that EnD^2 can approach the classification performance of an ensemble, and outperforms both standard DNNs and Ensemble Distillation on the tasks of misclassification and out-of-distribution input detection.

Deformable Kernels: Adapting Effective Receptive Fields for Object Deformation

- Hang Gao, Xizhou Zhu, Stephen Lin, Jifeng Dai
- abstract@open-review(Poster): Convolutional networks are not aware of an object's geometric variations, which leads to inefficient utilization of model and data capacity. To overcome this issue, recent works on deformation modeling seek to spatially reconfigure the data towards a common arrangement such that semantic recognition suffers less from deformation. This is typically done by augmenting static operators with learned free-form sampling grids in the image space, dynamically tuned to the data and task for adapting the receptive field. Yet adapting the receptive field does not quite reach the actual goal -- what really matters to the network is the *effective* receptive field (ERF), which reflects how much each pixel contributes. It is thus natural to design other approaches to adapt the ERF directly during runtime. In this work, we instantiate one possible solution as Deformable Kernels (DKs), a family of novel and generic convolutional operators for handling object deformations by directly adapting the ERF while leaving the receptive field untouched. At the heart of our method is the ability to resample the original kernel space towards recovering the deformation of objects. This approach is justified with theoretical insights that the ERF is strictly determined by data sampling locations and kernel values. We implement DKs as generic drop-in replacements of rigid kernels and conduct a series of empirical studies whose results conform with our theories. Over several tasks and standard base models, our approach compares favorably against prior works that adapt during runtime. In addition, further experiments suggest a working mechanism orthogonal and complementary to previous works.

VL-BERT: Pre-training of Generic Visual-Linguistic Representations

- Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, Jifeng Dai
- abstract@open-review(Poster): We introduce a new pre-trainable generic representation for visual-linguistic tasks, called Visual-Linguistic BERT (VL-BERT for short). VL-BERT adopts the simple yet powerful Transformer model as the backbone, and extends it to take both visual and linguistic embedded features as input. In it, each element of the input is either of a word from the input sentence, or a region-of-interest (RoI) from the input image. It is designed to fit for most of the visual-linguistic downstream tasks. To better exploit the generic representation, we pre-train VL-BERT on the massive-scale Conceptual Captions dataset, together with text-only corpus. Extensive empirical analysis demonstrates that the pre-training procedure can better align the visual-linguistic clues and benefit the downstream tasks, such as visual commonsense reasoning, visual question answering and referring expression comprehension. It is worth noting that VL-BERT achieved the first place of single model on the leaderboard of the VCR benchmark.

Optimistic Exploration even with a Pessimistic Initialisation

- Tabish Rashid, Bei Peng, Wendelin Boehmer, Shimon Whiteson
- abstract@open-review(Poster): Optimistic initialisation is an effective strategy for efficient exploration in reinforcement learning (RL). In the tabular case, all provably efficient model-free algorithms rely on it. However, model-free deep RL algorithms do not use optimistic initialisation despite taking inspiration from these provably efficient tabular algorithms. In particular, in scenarios with only positive rewards, Q-values are initialised at their lowest possible values due to commonly used network initialisation schemes, a pessimistic initialisation. Merely initialising the network to output optimistic Q-values is not enough, since we cannot ensure that they remain optimistic for novel state-action pairs, which is crucial for exploration. We propose a simple count-based augmentation to pessimistically initialised Q-values that separates the source of optimism from the neural network. We show that this scheme is provably efficient in the tabular setting and extend it to the deep RL setting. Our algorithm, Optimistic Pessimistically Initialised Q-Learning (OPIQ), augments the Q-value estimates of a DQN-based agent with count-derived bonuses to ensure optimism during both action selection and bootstrapping. We show that OPIQ outperforms non-optimistic DQN variants that utilise a pseudocount-based intrinsic motivation in hard exploration tasks, and that it predicts optimistic estimates for novel state-action pairs.

Certified Robustness for Top-k Predictions against Adversarial Perturbations via Randomized Smoothing

- Jinyuan Jia, Xiaoyu Cao, Binghui Wang, Neil Zhenqiang Gong
- abstract@open-review(Poster): It is well-known that classifiers are vulnerable to adversarial perturbations. To defend against adversarial perturbations, various certified robustness results have been derived. However, existing certified robustnesses are limited to top-1 predictions. In many real-world applications, top-k predictions are more relevant. In this work, we aim to derive certified robustness for top-k predictions. In particular, our certified robustness is based on randomized smoothing, which turns any classifier to a new classifier via adding noise to an input example. We adopt randomized smoothing because it is scalable to large-scale neural networks and applicable to any classifier. We derive a tight robustness in ℓ_2 norm for top-k predictions when using randomized smoothing with Gaussian noise. We find that generalizing the certified robustness from top-1 to top-k predictions faces significant technical challenges. We also empirically evaluate our method on CIFAR10 and ImageNet. For example, our method can obtain an ImageNet classifier with a certified top-5 accuracy of 62.8\% when the ℓ_2-norms of the adversarial perturbations are less than 0.5 (=127/255). Our code is publicly available at: \url{https://github.com/jjy1994/Certify_Topk}.

Identifying through Flows for Recovering Latent Representations

- Shen Li, Bryan Hooi, Gim Hee Lee
- abstract@open-review(Poster): Identifiability, or recovery of the true latent representations from which the observed data originates, is de facto a fundamental goal of representation learning. Yet, most deep generative models do not address the question of identifiability, and thus fail to deliver on the promise of the recovery of the true latent sources that generate the observations. Recent work proposed identifiable generative modelling using variational autoencoders (iVAE) with a theory of identifiability. Due to the intractablity of KL divergence between variational approximate posterior and the true posterior, however, iVAE has to maximize the evidence lower bound (ELBO) of the marginal likelihood, leading to suboptimal solutions in both theory and practice. In contrast, we propose an identifiable framework for estimating latent representations using a flow-based model (iFlow). Our approach directly maximizes the marginal likelihood, allowing for theoretical guarantees on identifiability, thereby dispensing with variational approximations. We derive its optimization objective in analytical form, making it possible to train iFlow in an end-to-end manner. Simulations on synthetic data validate the correctness and effectiveness of our proposed method and demonstrate its practical advantages over other existing methods.

Robust training with ensemble consensus

- Jisoo Lee, Sae-Young Chung
- abstract@open-review(Poster): Since deep neural networks are over-parameterized, they can memorize noisy examples. We address such a memorization issue in the presence of label noise. From the fact that deep neural networks cannot generalize to neighborhoods of memorized features, we hypothesize that noisy examples do not consistently incur small losses on the network under a certain perturbation. Based on this, we propose a novel training method called Learning with Ensemble Consensus (LEC) that prevents overfitting to noisy examples by removing them based on the consensus of an ensemble of perturbed networks. One of the proposed LECs, LTEC outperforms the current state-of-the-art methods on noisy MNIST, CIFAR-10, and CIFAR-100 in an efficient manner.

Self-Adversarial Learning with Comparative Discrimination for Text Generation

- Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, Ming Zhou
- abstract@open-review(Poster): Conventional Generative Adversarial Networks (GANs) for text generation tend to have issues of reward sparsity and mode collapse that affect the quality and diversity of generated samples. To address the issues, we propose a novel self-adversarial learning (SAL) paradigm for improving GANs' performance in text generation. In contrast to standard GANs that use a binary classifier as its discriminator to predict whether a sample is real or generated, SAL employs a comparative discriminator which is a pairwise classifier for comparing the text quality between a

pair of samples. During training, SAL rewards the generator when its currently generated sentence is found to be better than its previously generated samples. This self-improvement reward mechanism allows the model to receive credits more easily and avoid collapsing towards the limited number of real samples, which not only helps alleviate the reward sparsity issue but also reduces the risk of mode collapse. Experiments on text generation benchmark datasets show that our proposed approach substantially improves both the quality and the diversity, and yields more stable performance compared to the previous GANs for text generation.

Vid2Game: Controllable Characters Extracted from Real-World Videos

- Oran Gafni, Lior Wolf, Yaniv Taigman
- abstract@open-review(Poster): We extract a controllable model from a video of a person performing a certain activity. The model generates novel image sequences of that person, according to user-defined control signals, typically marking the displacement of the moving body. The generated video can have an arbitrary background, and effectively capture both the dynamics and appearance of the person.

The method is based on two networks. The first maps a current pose, and a single-instance control signal to the next pose. The second maps the current pose, the new pose, and a given background, to an output frame. Both networks include multiple novelties that enable high-quality performance. This is demonstrated on multiple characters extracted from various videos of dancers and athletes.

Action Semantics Network: Considering the Effects of Actions in Multiagent Systems

- Weixun Wang, Tianpei Yang, Yong Liu, Jianye Hao, Xiaotian Hao, Yujing Hu, Yingfeng Chen, Changjie Fan, Yang Gao
- abstract@open-review(Poster): In multiagent systems (MASs), each agent makes individual decisions but all of them contribute globally to the system evolution. Learning in MASs is difficult since each agent's selection of actions must take place in the presence of other co-learning agents. Moreover, the environmental stochasticity and uncertainties increase exponentially with the increase in the number of agents. Previous works borrow various multiagent coordination mechanisms into deep learning architecture to facilitate multiagent coordination. However, none of them explicitly consider action semantics between agents that different actions have different influences on other agents. In this paper, we propose a novel network architecture, named Action Semantics Network (ASN), that explicitly represents such action semantics between agents. ASN characterizes different actions' influence on other agents using neural networks based on the action semantics between them. ASN can be easily combined with existing deep reinforcement learning (DRL) algorithms to boost their performance. Experimental results on StarCraft II micromanagement and Neural MMO show ASN significantly improves the performance of state-of-the-art DRL approaches compared with several network architectures.

Learning Efficient Parameter Server Synchronization Policies for Distributed SGD

- Rong Zhu, Sheng Yang, Andreas Pfadler, Zhengping Qian, Jingren Zhou
- abstract@open-review(Poster): We apply a reinforcement learning (RL) based approach to learning optimal synchronization policies used for Parameter Server-based distributed training of machine learning models with Stochastic Gradient Descent (SGD). Utilizing a formal synchronization policy description in the PS-setting, we are able to derive a suitable and compact description of states and actions, allowing us to efficiently use the standard off-the-shelf deep Q-learning algorithm. As a result, we are able to learn synchronization policies which generalize to different cluster environments, different training datasets and small model variations and (most importantly) lead to considerable decreases in training time when compared to standard policies such as bulk synchronous parallel (BSP), asynchronous parallel (ASP), or stale synchronous parallel (SSP). To support our claims we present extensive numerical results obtained from experiments performed in simulated cluster environments. In our experiments training time is reduced by 44 on average and learned policies generalize to multiple unseen circumstances.

Relational State-Space Model for Stochastic Multi-Object Systems

- Fan Yang, Ling Chen, Fan Zhou, Yusong Gao, Wei Cao
- abstract@open-review(Poster): Real-world dynamical systems often consist of multiple stochastic subsystems that interact with each other. Modeling and forecasting the behavior of such dynamics are generally not easy, due to the inherent hardness in understanding the complicated interactions and evolutions of their constituents. This paper introduces the relational state-space model (R-SSM), a sequential hierarchical latent variable model that makes use of graph neural networks (GNNs) to simulate the joint state transitions of multiple correlated objects. By letting GNNs cooperate with SSM, R-SSM provides a flexible way to incorporate relational information into the modeling of multi-object dynamics. We further suggest augmenting the model with normalizing flows instantiated for vertex-indexed random variables and propose two auxiliary contrastive objectives to facilitate the learning. The utility of R-SSM is empirically evaluated on synthetic and real time series datasets.

Piecewise linear activations substantially shape the loss surfaces of neural networks

- Fengxiang He, Bohan Wang, Dacheng Tao
- abstract@open-review(Poster): Understanding the loss surface of a neural network is fundamentally important to the understanding of deep learning. This paper presents how piecewise linear activation functions substantially shape the loss surfaces of neural networks. We first prove that {\it the loss surfaces of many neural networks have infinite spurious local minima} which are defined as the local minima with higher empirical risks than the global minima. Our result demonstrates that the networks with piecewise linear activations possess substantial differences to the well-studied linear neural networks. This result holds for any neural network with arbitrary depth and arbitrary piecewise linear activation functions (excluding linear functions) under most loss functions in practice. Essentially, the underlying assumptions are consistent with most practical circumstances where the output layer is narrower than any hidden layer. In addition, the loss surface of a neural network with piecewise linear activations is partitioned into multiple smooth and multilinear cells by nondifferentiable boundaries. The constructed spurious local minima are concentrated in one cell as a valley: they are connected with each other by a continuous path, on which empirical risk is invariant. Further for one-hidden-layer networks, we prove that all local minima in a cell constitute an equivalence class; they are concentrated in a valley; and they are all global minima in the cell.

Novelty Detection Via Blurring

- Sungik Choi, Sae-Young Chung
- abstract@open-review(Poster): Conventional out-of-distribution (OOD) detection schemes based on variational autoencoder or Random Network Distillation (RND) are known to assign lower uncertainty to the OOD data than the target distribution. In this work, we discover that such conventional novelty detection schemes are also vulnerable to the blurred images. Based on the observation, we construct a novel RND-based OOD detector, SVD-RND, that utilizes blurred images during training. Our detector is simple, efficient in test time, and outperforms baseline OOD detectors in various domains. Further results show that SVD-RND learns a better target distribution representation than the baselines. Finally, SVD-RND combined with geometric transform achieves near-perfect detection accuracy in CelebA domain.

Bounds on Over-Parameterization for Guaranteed Existence of Descent Paths in Shallow ReLU Networks

- Arsalan Sharifnassab, Saber Salehkaleybar, S. Jamaloddin Golestani
- abstract@open-review(Poster): We study the landscape of squared loss in neural networks with one-hidden layer and ReLU activation functions. Let m and d be the widths of hidden and input layers, respectively. We show that there exist poor local minima with positive curvature for some training sets

of size $n\geq m+2d-2$. By positive curvature of a local minimum, we mean that within a small neighborhood the loss function is strictly increasing in all directions. Consequently, for such training sets, there are initialization of weights from which there is no descent path to global optima. It is known that for $n\le m$, there always exist descent paths to global optima from all initial weights. In this perspective, our results provide a somewhat sharp characterization of the over-parameterization required for "existence of descent paths" in the loss landscape.

Data-Independent Neural Pruning via Coresets

- Ben Mussay, Margarita Osadchy, Vladimir Braverman, Samson Zhou, Dan Feldman
- abstract@open-review(Poster): Previous work showed empirically that large neural networks can be significantly reduced in size while preserving their accuracy. Model compression became a central research topic, as it is crucial for deployment of neural networks on devices with limited computational and memory resources. The majority of the compression methods are based on heuristics and offer no worst-case guarantees on the trade-off between the compression rate and the approximation error for an arbitrarily new sample.

We propose the first efficient, data-independent neural pruning algorithm with a provable trade-off between its compression rate and the approximation error for any future test sample. Our method is based on the coreset framework, which finds a small weighted subset of points that provably approximates the original inputs. Specifically, we approximate the output of a layer of neurons by a coreset of neurons in the previous layer and discard the rest. We apply this framework in a layer-by-layer fashion from the top to the bottom. Unlike previous works, our coreset is data independent, meaning that it provably guarantees the accuracy of the function for any input $x\in \mathbb{R}^d$, including an adversarial one. We demonstrate the effectiveness of our method on popular network architectures. In particular, our coresets yield 90% compression of the LeNet-300-100 architecture on MNIST while improving the accuracy.

Deep Network Classification by Scattering and Homotopy Dictionary Learning

- John Zarka, Louis Thiry, Tomas Angles, Stephane Mallat
- abstract@open-review(Poster): We introduce a sparse scattering deep convolutional neural network, which provides a simple model to analyze properties of deep representation learning for classification. Learning a single dictionary matrix with a classifier yields a higher classification accuracy than AlexNet over the ImageNet 2012 dataset. The network first applies a scattering transform that linearizes variabilities due to geometric transformations such as translations and small deformations. A sparse ℓ^1 dictionary coding reduces intra-class variability while preserving class separation through projections over unions of linear spaces. It is implemented in a deep convolutional network with a homotopy algorithm having an exponential convergence. A convergence proof is given in a general framework that includes ALISTA. Classification results are analyzed on ImageNet.

Q-learning with UCB Exploration is Sample Efficient for Infinite-Horizon MDP

- Yuanhao Wang, Kefan Dong, Xiaoyu Chen, Liwei Wang
- abstract@open-review(Poster): A fundamental question in reinforcement learning is whether model-free algorithms are sample efficient. Recently, Jin et al. (2018) proposed a Q-learning algorithm with UCB exploration policy, and proved it has nearly optimal regret bound for finite-horizon episodic MDP. In this paper, we adapt Q-learning with UCB-exploration bonus to infinite-horizon MDP with discounted rewards \emph{without} accessing a generative model. We show that the \textit{sample complexity of exploration} of our algorithm is bounded by $\tilde{O}({\frac{SA}{\epsilon^2(1-\gamma)^7}})$. This improves the previously best known result of $\tilde{O}({\frac{SA}{\epsilon^4(1-\gamma)^8}})$ in this setting achieved by delayed Q-learning (Strehlet al., 2006),, and matches the lower bound in terms of ϵ as well as S and A up to logarithmic factors.

Mixout: Effective Regularization to Finetune Large-scale Pretrained Language Models

- Cheolhyoung Lee, Kyunghyun Cho, Wanmo Kang
- abstract@open-review(Poster): In natural language processing, it has been observed recently that generalization could be greatly improved by finetuning a large-scale language model pretrained on a large unlabeled corpus. Despite its recent success and wide adoption, finetuning a large pretrained language model on a downstream task is prone to degenerate performance when there are only a small number of training instances available. In this paper, we introduce a new regularization technique, to which we refer as “mixout”, motivated by dropout. Mixout stochastically mixes the parameters of two models. We show that our mixout technique regularizes learning to minimize the deviation from one of the two models and that the strength of regularization adapts along the optimization trajectory. We empirically evaluate the proposed mixout and its variants on finetuning a pretrained language model on downstream tasks. More specifically, we demonstrate that the stability of finetuning and the average accuracy greatly increase when we use the proposed approach to regularize finetuning of BERT on downstream tasks in GLUE.

I Am Going MAD: Maximum Discrepancy Competition for Comparing Classifiers Adaptively

- Haotao Wang, Tianlong Chen, Zhangyang Wang, Kede Ma
- abstract@open-review(Poster): The learning of hierarchical representations for image classification has experienced an impressive series of successes due in part to the availability of large-scale labeled data for training. On the other hand, the trained classifiers have traditionally been evaluated on small and fixed sets of test images, which are deemed to be extremely sparsely distributed in the space of all natural images. It is thus questionable whether recent performance improvements on the excessively re-used test sets generalize to real-world natural images with much richer content variations. Inspired by efficient stimulus selection for testing perceptual models in psychophysical and physiological studies, we present an alternative framework for comparing image classifiers, which we name the MAximum Discrepancy (MAD) competition. Rather than comparing image classifiers using fixed test images, we adaptively sample a small test set from an arbitrarily large corpus of unlabeled images so as to maximize the discrepancies between the classifiers, measured by the distance over WordNet hierarchy. Human labeling on the resulting model-dependent image sets reveals the relative performance of the competing classifiers, and provides useful insights on potential ways to improve them. We report the MAD competition results of eleven ImageNet classifiers while noting that the framework is readily extensible and cost-effective to add future classifiers into the competition. Codes can be found at https://github.com/TAMU-VITA/MAD.

Black-Box Adversarial Attack with Transferable Model-based Embedding

- Zhichao Huang, Tong Zhang
- abstract@open-review(Poster): We present a new method for black-box adversarial attack. Unlike previous methods that combined transfer-based and scored-based methods by using the gradient or initialization of a surrogate white-box model, this new method tries to learn a low-dimensional embedding using a pretrained model, and then performs efficient search within the embedding space to attack an unknown target network. The method produces adversarial perturbations with high level semantic patterns that are easily transferable. We show that this approach can greatly improve the query efficiency of black-box adversarial attack across different target network architectures. We evaluate our approach on MNIST, ImageNet and Google Cloud Vision API, resulting in a significant reduction on the number of queries. We also attack adversarially defended networks on CIFAR10 and ImageNet, where our method not only reduces the number of queries, but also improves the attack success rate.

Cross-Domain Few-Shot Classification via Learned Feature-Wise Transformation

- Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, Ming-Hsuan Yang

- abstract@open-review(Spotlight): Few-shot classification aims to recognize novel categories with only few labeled images in each class. Existing metric-based few-shot classification algorithms predict categories by comparing the feature embeddings of query images with those from a few labeled images (support examples) using a learned metric function. While promising performance has been demonstrated, these methods often fail to generalize to unseen domains due to large discrepancy of the feature distribution across domains. In this work, we address the problem of few-shot classification under domain shifts for metric-based methods. Our core idea is to use feature-wise transformation layers for augmenting the image features using affine transforms to simulate various feature distributions under different domains in the training stage. To capture variations of the feature distributions under different domains, we further apply a learning-to-learn approach to search for the hyper-parameters of the feature-wise transformation layers. We conduct extensive experiments and ablation studies under the domain generalization setting using five few-shot classification datasets: mini-ImageNet, CUB, Cars, Places, and Plantae. Experimental results demonstrate that the proposed feature-wise transformation layer is applicable to various metric-based models, and provides consistent improvements on the few-shot classification performance under domain shift.

Compositional languages emerge in a neural iterated learning model

- Yi Ren, Shangmin Guo, Matthieu Labeau, Shay B. Cohen, Simon Kirby
- abstract@open-review(Poster): The principle of compositionality, which enables natural language to represent complex concepts via a structured combination of simpler ones, allows us to convey an open-ended set of messages using a limited vocabulary. If compositionality is indeed a natural property of language, we may expect it to appear in communication protocols that are created by neural agents via grounded language learning. Inspired by the iterated learning framework, which simulates the process of language evolution, we propose an effective neural iterated learning algorithm that, when applied to interacting neural agents, facilitates the emergence of a more structured type of language. Indeed, these languages provide specific advantages to neural agents during training, which translates as a larger posterior probability, which is then incrementally amplified via the iterated learning procedure. Our experiments confirm our analysis, and also demonstrate that the emerged languages largely improve the generalization of the neural agent communication.

Population-Guided Parallel Policy Search for Reinforcement Learning

- Whiyoung Jung, Giseung Park, Youngchul Sung
- abstract@open-review(Poster): In this paper, a new population-guided parallel learning scheme is proposed to enhance the performance of off-policy reinforcement learning (RL). In the proposed scheme, multiple identical learners with their own value-functions and policies share a common experience replay buffer, and search a good policy in collaboration with the guidance of the best policy information. The key point is that the information of the best policy is fused in a soft manner by constructing an augmented loss function for policy update to enlarge the overall search region by the multiple learners. The guidance by the previous best policy and the enlarged range enable faster and better policy search, and monotone improvement of the expected cumulative return by the proposed scheme is proved theoretically. Working algorithms are constructed by applying the proposed scheme to the twin delayed deep deterministic (TD3) policy gradient algorithm, and numerical results show that the constructed P3S-TD3 outperforms most of the current state-of-the-art RL algorithms, and the gain is significant in the case of sparse reward environment.

Variational Recurrent Models for Solving Partially Observable Control Tasks

- Dongqi Han, Kenji Doya, Jun Tani
- abstract@open-review(Poster): In partially observable (PO) environments, deep reinforcement learning (RL) agents often suffer from unsatisfactory performance, since two problems need to be tackled together: how to extract information from the raw observations to solve the task, and how to improve the policy. In this study, we propose an RL algorithm for solving PO tasks. Our method comprises two parts: a variational recurrent model (VRM) for modeling the environment, and an RL controller that has access to both the environment and the VRM. The proposed algorithm was tested in two types of PO robotic control tasks, those in which either coordinates or velocities were not observable and those that require long-term memorization. Our experiments show that the proposed algorithm achieved better data efficiency and/or learned more optimal policy than other alternative approaches in tasks in which unobserved states cannot be inferred from raw observations in a simple manner.

GAT: Generative Adversarial Training for Adversarial Example Detection and Robust Classification

- Xuwang Yin, Soheil Kolouri, Gustavo K Rohde
- abstract@open-review(Poster): The vulnerabilities of deep neural networks against adversarial examples have become a significant concern for deploying these models in sensitive domains. Devising a definitive defense against such attacks is proven to be challenging, and the methods relying on detecting adversarial samples are only valid when the attacker is oblivious to the detection mechanism. In this paper we propose a principled adversarial example detection method that can withstand norm-constrained white-box attacks. Inspired by one-versus-the-rest classification, in a K class classification problem, we train K binary classifiers where the i-th binary classifier is used to distinguish between clean data of class i and adversarially perturbed samples of other classes. At test time, we first use a trained classifier to get the predicted label (say k) of the input, and then use the k-th binary classifier to determine whether the input is a clean sample (of class k) or an adversarially perturbed example (of other classes). We further devise a generative approach to detecting/classifying adversarial examples by interpreting each binary classifier as an unnormalized density model of the class-conditional data. We provide comprehensive evaluation of the above adversarial example detection/classification methods, and demonstrate their competitive performances and compelling properties. Code is available at https://github.com/xuwangyin/GAT-Generative-Adversarial-Training

Posterior sampling for multi-agent reinforcement learning: solving extensive games with imperfect information

- Yichi Zhou, Jialian Li, Jun Zhu
- abstract@open-review(Talk): Posterior sampling for reinforcement learning (PSRL) is a useful framework for making decisions in an unknown environment. PSRL maintains a posterior distribution of the environment and then makes planning on the environment sampled from the posterior distribution. Though PSRL works well on single-agent reinforcement learning problems, how to apply PSRL to multi-agent reinforcement learning problems is relatively unexplored. In this work, we extend PSRL to two-player zero-sum extensive-games with imperfect information (TEGI), which is a class of multi-agent systems. More specifically, we combine PSRL with counterfactual regret minimization (CFR), which is the leading algorithm for TEGI with a known environment. Our main contribution is a novel design of interaction strategies. With our interaction strategies, our algorithm provably converges to the Nash Equilibrium at a rate of $O(\sqrt{\log T/T})$. Empirical results show that our algorithm works well.

Detecting and Diagnosing Adversarial Images with Class-Conditional Capsule Reconstructions

- Yao Qin, Nicholas Frosst, Sara Sabour, Colin Raffel, Garrison Cottrell, Geoffrey Hinton
- abstract@open-review(Poster): Adversarial examples raise questions about whether neural network models are sensitive to the same visual features as humans. In this paper, we first detect adversarial examples or otherwise corrupted images based on a class-conditional reconstruction of the input. To specifically attack our detection mechanism, we propose the Reconstructive Attack which seeks both to cause a misclassification and a low reconstruction error. This reconstructive attack produces undetected adversarial examples but with much smaller success rate. Among all these attacks, we find that CapsNets always perform better than convolutional networks. Then, we diagnose the adversarial examples for CapsNets and find that the success of the reconstructive attack is highly related to the visual similarity between the source and target class. Additionally, the resulting perturbations can cause the input image to appear visually more like the target class and hence become non-adversarial. This suggests that CapsNets use features that are more aligned with human perception and have the potential to address the central issue raised by adversarial examples.

MACER: Attack-free and Scalable Robust Training via Maximizing Certified Radius

- Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, Liwei Wang
- abstract@open-review(Poster): Adversarial training is one of the most popular ways to learn robust models but is usually attack-dependent and time costly. In this paper, we propose the MACER algorithm, which learns robust models without using adversarial training but performs better than all existing provable l2-defenses. Recent work shows that randomized smoothing can be used to provide a certified l2 radius to smoothed classifiers, and our algorithm trains provably robust smoothed classifiers via MAximizing the CErtified Radius (MACER). The attack-free characteristic makes MACER faster to train and easier to optimize. In our experiments, we show that our method can be applied to modern deep neural networks on a wide range of datasets, including Cifar-10, ImageNet, MNIST, and SVHN. For all tasks, MACER spends less training time than state-of-the-art adversarial training algorithms, and the learned models achieve larger average certified radius.

Semantically-Guided Representation Learning for Self-Supervised Monocular Depth

- Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, Adrien Gaidon
- abstract@open-review(Poster): Self-supervised learning is showing great promise for monocular depth estimation, using geometry as the only source of supervision. Depth networks are indeed capable of learning representations that relate visual appearance to 3D properties by implicitly leveraging category-level patterns. In this work we investigate how to leverage more directly this semantic structure to guide geometric representation learning, while remaining in the self-supervised regime. Instead of using semantic labels and proxy losses in a multi-task approach, we propose a new architecture leveraging fixed pretrained semantic segmentation networks to guide self-supervised representation learning via pixel-adaptive convolutions. Furthermore, we propose a two-stage training process to overcome a common semantic bias on dynamic objects via resampling. Our method improves upon the state of the art for self-supervised monocular depth prediction over all pixels, fine-grained details, and per semantic categories.

Stochastic AUC Maximization with Deep Neural Networks

- Mingrui Liu, Zhuoning Yuan, Yiming Ying, Tianbao Yang
- abstract@open-review(Poster): Stochastic AUC maximization has garnered an increasing interest due to better fit to imbalanced data classification. However, existing works are limited to stochastic AUC maximization with a linear predictive model, which restricts its predictive power when dealing with extremely complex data. In this paper, we consider stochastic AUC maximization problem with a deep neural network as the predictive model. Building on the saddle point reformulation of a surrogated loss of AUC, the problem can be cast into a {\it non-convex concave} min-max problem. The main contribution made in this paper is to make stochastic AUC maximization more practical for deep neural networks and big data with theoretical insights as well. In particular, we propose to explore Polyak-\L{}ojasiewicz (PL) condition that has been proved and observed in deep learning, which enables us to develop new stochastic algorithms with even faster convergence rate and more practical step size scheme. An AdaGrad-style algorithm is also analyzed under the PL condition with adaptive convergence rate. Our experimental results demonstrate the effectiveness of the proposed algorithms.

Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity

- Jingzhao Zhang, Tianxing He, Suvrit Sra, Ali Jadbabaie
- abstract@open-review(Talk): We provide a theoretical explanation for the effectiveness of gradient clipping in training deep neural networks. The key ingredient is a new smoothness condition derived from practical neural network training examples. We observe that gradient smoothness, a concept central to the analysis of first-order optimization algorithms that is often assumed to be a constant, demonstrates significant variability along the training trajectory of deep neural networks. Further, this smoothness positively correlates with the gradient norm, and contrary to standard assumptions in the literature, it can grow with the norm of the gradient. These empirical observations limit the applicability of existing theoretical analyses of algorithms that rely on a fixed bound on smoothness. These observations motivate us to introduce a novel relaxation of gradient smoothness that is weaker than the commonly used Lipschitz smoothness assumption. Under the new condition, we prove that two popular methods, namely, gradient clipping and normalized gradient, converge arbitrarily faster than gradient descent with fixed stepsize. We further explain why such adaptively scaled gradient methods can accelerate empirical convergence and verify our results empirically in popular neural network training settings.

Difference-Seeking Generative Adversarial Network--Unseen Sample Generation

- Yi Lin Sung, Sung-Hsien Hsieh, Soo-Chang Pei, Chun-Shien Lu
- abstract@open-review(Poster): Unseen data, which are not samples from the distribution of training data and are difficult to collect, have exhibited importance in numerous applications, ({\em e.g.,} novelty detection, semi-supervised learning, and adversarial training). In this paper, we introduce a general framework called \textbf{d}ifference-\textbf{s}eeking \textbf{g}enerative \textbf{a}dversarial \textbf{n}etwork (DSGAN), to generate various types of unseen data. Its novelty is the consideration of the probability density of the unseen data distribution as the difference between two distributions $p_{\bar{d}}$ and p_{d} whose samples are relatively easy to collect.

The DSGAN can learn the target distribution, p_{t}, (or the unseen data distribution) from only the samples from the two distributions, p_{d} and $p_{\bar{d}}$. In our scenario, p_d is the distribution of the seen data, and $p_{\bar{d}}$ can be obtained from p_{d} via simple operations, so that we only need the samples of p_{d} during the training. Two key applications, semi-supervised learning and novelty detection, are taken as case studies to illustrate that the DSGAN enables the production of various unseen data. We also provide theoretical analyses about the convergence of the DSGAN.

FasterSeg: Searching for Faster Real-time Semantic Segmentation

- Wuyang Chen, Xinyu Gong, Xianming Liu, Qian Zhang, Yuan Li, Zhangyang Wang
- abstract@open-review(Poster): We present FasterSeg, an automatically designed semantic segmentation network with not only state-of-the-art performance but also faster speed than current methods. Utilizing neural architecture search (NAS), FasterSeg is discovered from a novel and broader search space integrating multi-resolution branches, that has been recently found to be vital in manually designed segmentation models. To better calibrate the balance between the goals of high accuracy and low latency, we propose a decoupled and fine-grained latency regularization, that effectively overcomes our observed phenomenons that the searched networks are prone to "collapsing" to low-latency yet poor-accuracy models. Moreover, we seamlessly extend FasterSeg to a new collaborative search (co-searching) framework, simultaneously searching for a teacher and a student network in the same single run. The teacher-student distillation further boosts the student model’s accuracy. Experiments on popular segmentation benchmarks demonstrate the competency of FasterSeg. For example, FasterSeg can run over 30% faster than the closest manually designed competitor on Cityscapes, while maintaining comparable accuracy.

LEARNING EXECUTION THROUGH NEURAL CODE FUSION

- Zhan Shi, Kevin Swersky, Daniel Tarlow, Parthasarathy Ranganathan, Milad Hashemi
- abstract@open-review(Poster): As the performance of computer systems stagnates due to the end of Moore’s Law, there is a need for new models that can understand and optimize the execution of general purpose code. While there is a growing body of work on using Graph Neural Networks (GNNs) to learn static representations of source code, these representations do not understand how code executes at runtime. In this work, we propose a new approach using GNNs to learn fused representations of general source code and its execution. Our approach defines a multi-task GNN over low-level representations of source code and program state (i.e., assembly code and dynamic memory states), converting complex source code constructs and data

structures into a simpler, more uniform format. We show that this leads to improved performance over similar methods that do not use execution and it opens the door to applying GNN models to new tasks that would not be feasible from static code alone. As an illustration of this, we apply the new model to challenging dynamic tasks (branch prediction and prefetching) from the SPEC CPU benchmark suite, outperforming the state-of-the-art by 26% and 45% respectively. Moreover, we use the learned fused graph embeddings to demonstrate transfer learning with high performance on an indirectly related algorithm classification task.

Editable Neural Networks

- Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitry Pyrkin, Sergei Popov, Artem Babenko
- abstract@open-review(Poster): These days deep neural networks are ubiquitously used in a wide range of tasks, from image classification and machine translation to face identification and self-driving cars. In many applications, a single model error can lead to devastating financial, reputational and even life-threatening consequences. Therefore, it is crucially important to correct model mistakes quickly as they appear. In this work, we investigate the problem of neural network editing - how one can efficiently patch a mistake of the model on a particular sample, without influencing the model behavior on other samples. Namely, we propose Editable Training, a model-agnostic training technique that encourages fast editing of the trained model. We empirically demonstrate the effectiveness of this method on large-scale image classification and machine translation tasks.

Can gradient clipping mitigate label noise?

- Aditya Krishna Menon, Ankit Singh Rawat, Sashank J. Reddi, Sanjiv Kumar
- abstract@open-review(Poster): Gradient clipping is a widely-used technique in the training of deep networks, and is generally motivated from an optimisation lens: informally, it controls the dynamics of iterates, thus enhancing the rate of convergence to a local minimum. This intuition has been made precise in a line of recent works, which show that suitable clipping can yield significantly faster convergence than vanilla gradient descent. In this paper, we propose a new lens for studying gradient clipping, namely, robustness: informally, one expects clipping to provide robustness to noise, since one does not overly trust any single sample. Surprisingly, we prove that for the common problem of label noise in classification, standard gradient clipping does not in general provide robustness. On the other hand, we show that a simple variant of gradient clipping is provably robust, and corresponds to suitably modifying the underlying loss function. This yields a simple, noise-robust alternative to the standard cross-entropy loss which performs well empirically.

Pretrained Encyclopedia: Weakly Supervised Knowledge-Pretrained Language Model

- Wenhan Xiong, Jingfei Du, William Yang Wang, Veselin Stoyanov
- abstract@open-review(Poster): Recent breakthroughs of pretrained language models have shown the effectiveness of self-supervised learning for a wide range of natural language processing (NLP) tasks. In addition to standard syntactic and semantic NLP tasks, pretrained models achieve strong improvements on tasks that involve real-world knowledge, suggesting that large-scale language modeling could be an implicit method to capture knowledge. In this work, we further investigate the extent to which pretrained models such as BERT capture knowledge using a zero-shot fact completion task. Moreover, we propose a simple yet effective weakly supervised pretraining objective, which explicitly forces the model to incorporate knowledge about real-world entities. Models trained with our new objective yield significant improvements on the fact completion task. When applied to downstream tasks, our model consistently outperforms BERT on four entity-related question answering datasets (i.e., WebQuestions, TriviaQA, SearchQA and Quasar-T) with an average 2.7 F1 improvements and a standard fine-grained entity typing dataset (i.e., FIGER) with 5.7 accuracy gains.

Pruned Graph Scattering Transforms

- Vassilis N. Ioannidis, Siheng Chen, Georgios B. Giannakis
- abstract@open-review(Poster): Graph convolutional networks (GCNs) have achieved remarkable performance in a variety of network science learning tasks. However, theoretical analysis of such approaches is still at its infancy. Graph scattering transforms (GSTs) are non-trainable deep GCN models that are amenable to generalization and stability analyses. The present work addresses some limitations of GSTs by introducing a novel so-termed pruned (p)GST approach. The resultant pruning algorithm is guided by a graph-spectrum-inspired criterion, and retains informative scattering features on-the-fly while bypassing the exponential complexity associated with GSTs. It is further established that pGSTs are stable to perturbations of the input graph signals with bounded energy. Experiments showcase that i) pGST performs comparably to the baseline GST that uses all scattering features, while achieving significant computational savings; ii) pGST achieves comparable performance to state-of-the-art GCNs; and iii) Graph data from various domains lead to different scattering patterns, suggesting domain-adaptive pGST network architectures.

DDSP: Differentiable Digital Signal Processing

- Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, Adam Roberts
- abstract@open-review(Spotlight): Most generative models of audio directly generate samples in one of two domains: time or frequency. While sufficient to express any signal, these representations are inefficient, as they do not utilize existing knowledge of how sound is generated and perceived. A third approach (vocoders/synthesizers) successfully incorporates strong domain knowledge of signal processing and perception, but has been less actively researched due to limited expressivity and difficulty integrating with modern auto-differentiation-based machine learning methods. In this paper, we introduce the Differentiable Digital Signal Processing (DDSP) library, which enables direct integration of classic signal processing elements with deep learning methods. Focusing on audio synthesis, we achieve high-fidelity generation without the need for large autoregressive models or adversarial losses, demonstrating that DDSP enables utilizing strong inductive biases without losing the expressive power of neural networks. Further, we show that combining interpretable modules permits manipulation of each separate model component, with applications such as independent control of pitch and loudness, realistic extrapolation to pitches not seen during training, blind dereverberation of room acoustics, transfer of extracted room acoustics to new environments, and transformation of timbre between disparate sources. In short, DDSP enables an interpretable and modular approach to generative modeling, without sacrificing the benefits of deep learning. The library will is available at https://github.com/magenta/ddsp and we encourage further contributions from the community and domain experts.

GLAD: Learning Sparse Graph Recovery

- Harsh Shrivastava, Xinshi Chen, Binghong Chen, Guanghui Lan, Srinivas Aluru, Han Liu, Le Song
- abstract@open-review(Poster): Recovering sparse conditional independence graphs from data is a fundamental problem in machine learning with wide applications. A popular formulation of the problem is an ℓ_1 regularized maximum likelihood estimation. Many convex optimization algorithms have been designed to solve this formulation to recover the graph structure. Recently, there is a surge of interest to learn algorithms directly based on data, and in this case, learn to map empirical covariance to the sparse precision matrix. However, it is a challenging task in this case, since the symmetric positive definiteness (SPD) and sparsity of the matrix are not easy to enforce in learned algorithms, and a direct mapping from data to precision matrix may contain many parameters. We propose a deep learning architecture, GLAD, which uses an Alternating Minimization (AM) algorithm as our model inductive bias, and learns the model parameters via supervised learning. We show that GLAD learns a very compact and effective model for recovering sparse graphs from data.

Neural Network Branching for Neural Network Verification

- Jingyue Lu, M. Pawan Kumar
- abstract@open-review(Talk): Formal verification of neural networks is essential for their deployment in safety-critical areas. Many available formal verification methods have been shown to be instances of a unified Branch and Bound (BaB) formulation. We propose a novel framework for designing an effective branching strategy for BaB. Specifically, we learn a graph neural network (GNN) to imitate the strong branching heuristic behaviour. Our framework differs from previous methods for learning to branch in two main aspects. Firstly, our framework directly treats the neural network we want to verify as a graph input for the GNN. Secondly, we develop an intuitive forward and backward embedding update schedule. Empirically, our framework achieves roughly 50% reduction in both the number of branches and the time required for verification on various convolutional networks when compared to the best available hand-designed branching strategy. In addition, we show that our GNN model enjoys both horizontal and vertical transferability. Horizontally, the model trained on easy properties performs well on properties of increased difficulty levels. Vertically, the model trained on small neural networks achieves similar performance on large neural networks.

VideoFlow: A Conditional Flow-Based Model for Stochastic Video Generation

- Manoj Kumar, Mohammad Babaeizadeh, Dumitru Erhan, Chelsea Finn, Sergey Levine, Laurent Dinh, Durk Kingma
- abstract@open-review(Poster): Generative models that can model and predict sequences of future events can, in principle, learn to capture complex real-world phenomena, such as physical interactions. However, a central challenge in video prediction is that the future is highly uncertain: a sequence of past observations of events can imply many possible futures. Although a number of recent works have studied probabilistic models that can represent uncertain futures, such models are either extremely expensive computationally as in the case of pixel-level autoregressive models, or do not directly optimize the likelihood of the data. To our knowledge, our work is the first to propose multi-frame video prediction with normalizing flows, which allows for direct optimization of the data likelihood, and produces high-quality stochastic predictions. We describe an approach for modeling the latent space dynamics, and demonstrate that flow-based generative models offer a viable and competitive approach to generative modeling of video.

Adversarial Policies: Attacking Deep Reinforcement Learning

- Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, Stuart Russell
- abstract@open-review(Poster): Deep reinforcement learning (RL) policies are known to be vulnerable to adversarial perturbations to their observations, similar to adversarial examples for classifiers. However, an attacker is not usually able to directly modify another agent's observations. This might lead one to wonder: is it possible to attack an RL agent simply by choosing an adversarial policy acting in a multi-agent environment so as to create natural observations that are adversarial? We demonstrate the existence of adversarial policies in zero-sum games between simulated humanoid robots with proprioceptive observations, against state-of-the-art victims trained via self-play to be robust to opponents. The adversarial policies reliably win against the victims but generate seemingly random and uncoordinated behavior. We find that these policies are more successful in high-dimensional environments, and induce substantially different activations in the victim policy network than when the victim plays against a normal opponent. Videos are available at https://adversarialpolicies.github.io/.

Escaping Saddle Points Faster with Stochastic Momentum

- Jun-Kun Wang, Chi-Heng Lin, Jacob Abernethy
- abstract@open-review(Poster): Stochastic gradient descent (SGD) with stochastic momentum is popular in nonconvex stochastic optimization and particularly for the training of deep neural networks. In standard SGD, parameters are updated by improving along the path of the gradient at the current iterate on a batch of examples, where the addition of a ``momentum'' term biases the update in the direction of the previous change in parameters. In non-stochastic convex optimization one can show that a momentum adjustment provably reduces convergence time in many settings, yet such results have been elusive in the stochastic and non-convex settings. At the same time, a widely-observed empirical phenomenon is that in training deep networks stochastic momentum appears to significantly improve convergence time, variants of it have flourished in the development of other popular update methods, e.g. ADAM, AMSGrad, etc. Yet theoretical justification for the use of stochastic momentum has remained a significant open question. In this paper we propose an answer: stochastic momentum improves deep network training because it modifies SGD to escape saddle points faster and, consequently, to more quickly find a second order stationary point. Our theoretical results also shed light on the related question of how to choose the ideal momentum parameter--our analysis suggests that $\beta \in [0,1)$ should be large (close to 1), which comports with empirical findings. We also provide experimental findings that further validate these conclusions.

Few-shot Text Classification with Distributional Signatures

- Yujia Bao, Menghua Wu, Shiyu Chang, Regina Barzilay
- abstract@open-review(Poster): In this paper, we explore meta-learning for few-shot text classification. Meta-learning has shown strong performance in computer vision, where low-level patterns are transferable across learning tasks. However, directly applying this approach to text is challenging--lexical features highly informative for one task may be insignificant for another. Thus, rather than learning solely from words, our model also leverages their distributional signatures, which encode pertinent word occurrence patterns. Our model is trained within a meta-learning framework to map these signatures into attention scores, which are then used to weight the lexical representations of words. We demonstrate that our model consistently outperforms prototypical networks learned on lexical knowledge (Snell et al., 2017) in both few-shot text classification and relation classification by a significant margin across six benchmark datasets (20.0% on average in 1-shot classification).

Geometric Insights into the Convergence of Nonlinear TD Learning

- David Brandfonbrener, Joan Bruna
- abstract@open-review(Poster): While there are convergence guarantees for temporal difference (TD) learning when using linear function approximators, the situation for nonlinear models is far less understood, and divergent examples are known. Here we take a first step towards extending theoretical convergence guarantees to TD learning with nonlinear function approximation. More precisely, we consider the expected learning dynamics of the TD(0) algorithm for value estimation. As the step-size converges to zero, these dynamics are defined by a nonlinear ODE which depends on the geometry of the space of function approximators, the structure of the underlying Markov chain, and their interaction. We find a set of function approximators that includes ReLU networks and has geometry amenable to TD learning regardless of environment, so that the solution performs about as well as linear TD in the worst case. Then, we show how environments that are more reversible induce dynamics that are better for TD learning and prove global convergence to the true value function for well-conditioned function approximators. Finally, we generalize a divergent counterexample to a family of divergent problems to demonstrate how the interaction between approximator and environment can go wrong and to motivate the assumptions needed to prove convergence.

Learning Self-Correctable Policies and Value Functions from Demonstrations with Negative Sampling

- Yuping Luo, Huazhe Xu, Tengyu Ma
- abstract@open-review(Poster): Imitation learning, followed by reinforcement learning algorithms, is a promising paradigm to solve complex control tasks sample-efficiently. However, learning from demonstrations often suffers from the covariate shift problem, which results in cascading errors of the learned policy. We introduce a notion of conservatively extrapolated value functions, which provably lead to policies with self-correction. We design an algorithm Value Iteration with Negative Sampling (VINS) that practically learns such value functions with conservative extrapolation. We show that VINS can correct mistakes of the behavioral cloning policy on simulated robotics benchmark tasks. We also propose the algorithm of using VINS to initialize a reinforcement learning algorithm, which is shown to outperform prior works in sample efficiency.

Exploring Model-based Planning with Policy Networks

- Tingwu Wang, Jimmy Ba
- abstract@open-review(Poster): Model-based reinforcement learning (MBRL) with model-predictive control or online planning has shown great potential for locomotion control tasks in both sample efficiency and asymptotic performance. Despite the successes, the existing planning methods search from candidate sequences randomly generated in the action space, which is inefficient in complex high-dimensional environments. In this paper, we propose a novel MBRL algorithm, model-based policy planning (POPLIN), that combines policy networks with online planning. More specifically, we formulate action planning at each time-step as an optimization problem using neural networks. We experiment with both optimization w.r.t. the action sequences initialized from the policy network, and also online optimization directly w.r.t. the parameters of the policy network. We show that POPLIN obtains state-of-the-art performance in the MuJoCo benchmarking environments, being about 3x more sample efficient than the state-of-the-art algorithms, such as PETS, TD3 and SAC. To explain the effectiveness of our algorithm, we show that the optimization surface in parameter space is smoother than in action space. Further more, we found the distilled policy network can be effectively applied without the expansive model predictive control during test time for some environments such as Cheetah. Code is released.

On Identifiability in Transformers

- Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, Roger Wattenhofer
- abstract@open-review(Poster): In this paper we delve deep in the Transformer architecture by investigating two of its core components: self-attention and contextual embeddings. In particular, we study the identifiability of attention weights and token embeddings, and the aggregation of context into hidden tokens. We show that, for sequences longer than the attention head dimension, attention weights are not identifiable. We propose effective attention as a complementary tool for improving explanatory interpretations based on attention. Furthermore, we show that input tokens retain to a large degree their identity across the model. We also find evidence suggesting that identity information is mainly encoded in the angle of the embeddings and gradually decreases with depth. Finally, we demonstrate strong mixing of input information in the generation of contextual embeddings by means of a novel quantification method based on gradient attribution. Overall, we show that self-attention distributions are not directly interpretable and present tools to better understand and further investigate Transformer models.

Automated curriculum generation through setter-solver interactions

- Sebastien Racaniere, Andrew Lampinen, Adam Santoro, David Reichert, Vlad Firoiu, Timothy Lillicrap
- abstract@open-review(Poster): Reinforcement learning algorithms use correlations between policies and rewards to improve agent performance. But in dynamic or sparsely rewarding environments these correlations are often too small, or rewarding events are too infrequent to make learning feasible. Human education instead relies on curricula –the breakdown of tasks into simpler, static challenges with dense rewards– to build up to complex behaviors. While curricula are also useful for artificial agents, hand-crafting them is time consuming. This has lead researchers to explore automatic curriculum generation. Here we explore automatic curriculum generation in rich,dynamic environments. Using a setter-solver paradigm we show the importance of considering goal validity, goal feasibility, and goal coverage to construct useful curricula. We demonstrate the success of our approach in rich but sparsely rewarding 2D and 3D environments, where an agent is tasked to achieve a single goal selected from a set of possible goals that varies between episodes, and identify challenges for future work. Finally, we demonstrate the value of a novel technique that guides agents towards a desired goal distribution. Altogether, these results represent a substantial step towards applying automatic task curricula to learn complex, otherwise unlearnable goals, and to our knowledge are the first to demonstrate automated curriculum generation for goal-conditioned agents in environments where the possible goals vary between episodes.

Progressive Memory Banks for Incremental Domain Adaptation

- Nabiha Asghar, Lili Mou, Kira A. Selby, Kevin D. Pantasdo, Pascal Poupart, Xin Jiang
- abstract@open-review(Poster): This paper addresses the problem of incremental domain adaptation (IDA) in natural language processing (NLP). We assume each domain comes one after another, and that we could only access data in the current domain. The goal of IDA is to build a unified model performing well on all the domains that we have encountered. We adopt the recurrent neural network (RNN) widely used in NLP, but augment it with a directly parameterized memory bank, which is retrieved by an attention mechanism at each step of RNN transition. The memory bank provides a natural way of IDA: when adapting our model to a new domain, we progressively add new slots to the memory bank, which increases the number of parameters, and thus the model capacity. We learn the new memory slots and fine-tune existing parameters by back-propagation. Experimental results show that our approach achieves significantly better performance than fine-tuning alone. Compared with expanding hidden states, our approach is more robust for old domains, shown by both empirical and theoretical results. Our model also outperforms previous work of IDA including elastic weight consolidation and progressive neural networks in the experiments.

What graph neural networks cannot learn: depth vs width

- Andreas Loukas
- abstract@open-review(Poster): This paper studies the expressive power of graph neural networks falling within the message-passing framework (GNNmp). Two results are presented. First, GNNmp are shown to be Turing universal under sufficient conditions on their depth, width, node attributes, and layer expressiveness. Second, it is discovered that GNNmp can lose a significant portion of their power when their depth and width is restricted. The proposed impossibility statements stem from a new technique that enables the repurposing of seminal results from distributed computing and leads to lower bounds for an array of decision, optimization, and estimation problems involving graphs. Strikingly, several of these problems are deemed impossible unless the product of a GNNmp's depth and width exceeds a polynomial of the graph size; this dependence remains significant even for tasks that appear simple or when considering approximation.

RTFM: Generalising to New Environment Dynamics via Reading

- Victor Zhong, Tim Rocktäschel, Edward Grefenstette
- abstract@open-review(Poster): Obtaining policies that can generalise to new environments in reinforcement learning is challenging. In this work, we demonstrate that language understanding via a reading policy learner is a promising vehicle for generalisation to new environments. We propose a grounded policy learning problem, Read to Fight Monsters (RTFM), in which the agent must jointly reason over a language goal, relevant dynamics described in a document, and environment observations. We procedurally generate environment dynamics and corresponding language descriptions of the dynamics, such that agents must read to understand new environment dynamics instead of memorising any particular information. In addition, we propose txt2π, a model that captures three-way interactions between the goal, document, and observations. On RTFM, txt2π generalises to new environments with dynamics not seen during training via reading. Furthermore, our model outperforms baselines such as FiLM and language-conditioned CNNs on RTFM. Through curriculum learning, txt2π produces policies that excel on complex RTFM tasks requiring several reasoning and coreference steps.

Input Complexity and Out-of-distribution Detection with Likelihood-based Generative Models

- Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, Jordi Luque

- abstract@open-review(Poster): Likelihood-based generative models are a promising resource to detect out-of-distribution (OOD) inputs which could compromise the robustness or reliability of a machine learning system. However, likelihoods derived from such models have been shown to be problematic for detecting certain types of inputs that significantly differ from training data. In this paper, we pose that this problem is due to the excessive influence that input complexity has in generative models' likelihoods. We report a set of experiments supporting this hypothesis, and use an estimate of input complexity to derive an efficient and parameter-free OOD score, which can be seen as a likelihood-ratio, akin to Bayesian model comparison. We find such score to perform comparably to, or even better than, existing OOD detection approaches under a wide range of data sets, models, model sizes, and complexity estimates.

Encoding word order in complex embeddings

- Benyou Wang, Donghao Zhao, Christina Lioma, Qiuchi Li, Peng Zhang, Jakob Grue Simonsen
- abstract@open-review(Spotlight): Sequential word order is important when processing text. Currently, neural networks (NNs) address this by modeling word position using position embeddings. The problem is that position embeddings capture the position of individual words, but not the ordered relationship (e.g., adjacency or precedence) between individual word positions. We present a novel and principled solution for modeling both the global absolute positions of words and their order relationships. Our solution generalizes word embeddings, previously defined as independent vectors, to continuous word functions over a variable (position). The benefit of continuous functions over variable positions is that word representations shift smoothly with increasing positions. Hence, word representations in different positions can correlate with each other in a continuous function. The general solution of these functions can be extended to complex-valued variants. We extend CNN, RNN and Transformer NNs to complex-valued versions to incorporate our complex embedding (we make all code available). Experiments on text classification, machine translation and language modeling show gains over both classical word embeddings and position-enriched word embeddings. To our knowledge, this is the first work in NLP to link imaginary numbers in complex-valued representations to concrete meanings (i.e., word order).

Functional vs. parametric equivalence of ReLU networks

- Mary Phuong, Christoph H. Lampert
- abstract@open-review(Poster): We address the following question: How redundant is the parameterisation of ReLU networks? Specifically, we consider transformations of the weight space which leave the function implemented by the network intact. Two such transformations are known for feed-forward architectures: permutation of neurons within a layer, and positive scaling of all incoming weights of a neuron coupled with inverse scaling of its outgoing weights. In this work, we show for architectures with non-increasing widths that permutation and scaling are in fact the only function-preserving weight transformations. For any eligible architecture we give an explicit construction of a neural network such that any other network that implements the same function can be obtained from the original one by the application of permutations and rescaling. The proof relies on a geometric understanding of boundaries between linear regions of ReLU networks, and we hope the developed mathematical tools are of independent interest.

Contrastive Learning of Structured World Models

- Thomas Kipf, Elise van der Pol, Max Welling
- abstract@open-review(Talk): A structured understanding of our world in terms of objects, relations, and hierarchies is an important component of human cognition. Learning such a structured world model from raw sensory data remains a challenge. As a step towards this goal, we introduce Contrastively-trained Structured World Models (C-SWMs). C-SWMs utilize a contrastive approach for representation learning in environments with compositional structure. We structure each state embedding as a set of object representations and their relations, modeled by a graph neural network. This allows objects to be discovered from raw pixel observations without direct supervision as part of the learning process. We evaluate C-SWMs on compositional environments involving multiple interacting objects that can be manipulated independently by an agent, simple Atari games, and a multi-object physics simulation. Our experiments demonstrate that C-SWMs can overcome limitations of models based on pixel reconstruction and outperform typical representatives of this model class in highly structured environments, while learning interpretable object-based representations.

Disentangling Factors of Variations Using Few Labels

- Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem
- abstract@open-review(Poster): Learning disentangled representations is considered a cornerstone problem in representation learning. Recently, Locatello et al. (2019) demonstrated that unsupervised disentanglement learning without inductive biases is theoretically impossible and that existing inductive biases and unsupervised methods do not allow to consistently learn disentangled representations. However, in many practical settings, one might have access to a limited amount of supervision, for example through manual labeling of (some) factors of variation in a few training examples. In this paper, we investigate the impact of such supervision on state-of-the-art disentanglement methods and perform a large scale study, training over 52000 models under well-defined and reproducible experimental conditions. We observe that a small number of labeled examples (0.01--0.5% of the data set), with potentially imprecise and incomplete labels, is sufficient to perform model selection on state-of-the-art unsupervised models. Further, we investigate the benefit of incorporating supervision into the training process. Overall, we empirically validate that with little and imprecise supervision it is possible to reliably learn disentangled representations.

A critical analysis of self-supervision, or what we can learn from a single image

- Asano YM., Rupprecht C., Vedaldi A.
- abstract@open-review(Poster): We look critically at popular self-supervision techniques for learning deep convolutional neural networks without manual labels. We show that three different and representative methods, BiGAN, RotNet and DeepCluster, can learn the first few layers of a convolutional network from a single image as well as using millions of images and manual labels, provided that strong data augmentation is used. However, for deeper layers the gap with manual supervision cannot be closed even if millions of unlabelled images are used for training. We conclude that: (1) the weights of the early layers of deep networks contain limited information about the statistics of natural images, that (2) such low-level statistics can be learned through self-supervision just as well as through strong supervision, and that (3) the low-level statistics can be captured via synthetic transformations instead of using a large image dataset.

Accelerating SGD with momentum for over-parameterized learning

- Chaoyue Liu, Mikhail Belkin
- abstract@open-review(Poster): Nesterov SGD is widely used for training modern neural networks and other machine learning models. Yet, its advantages over SGD have not been theoretically clarified. Indeed, as we show in this paper, both theoretically and empirically, Nesterov SGD with any parameter selection does not in general provide acceleration over ordinary SGD. Furthermore, Nesterov SGD may diverge for step sizes that ensure convergence of ordinary SGD. This is in contrast to the classical results in the deterministic setting, where the same step size ensures accelerated convergence of the Nesterov's method over optimal gradient descent.

To address the non-acceleration issue, we introduce a compensation term to Nesterov SGD. The resulting algorithm, which we call MaSS, converges for same step sizes as SGD. We prove that MaSS obtains an accelerated convergence rates over SGD for any mini-batch size in the linear setting. For full batch, the convergence rate of MaSS matches the well-known accelerated rate of the Nesterov's method.

We also analyze the practically important question of the dependence of the convergence rate and optimal hyper-parameters on the mini-batch size, demonstrating three distinct regimes: linear scaling, diminishing returns and saturation.

Experimental evaluation of MaSS for several standard architectures of deep networks, including ResNet and convolutional networks, shows improved performance over SGD, Nesterov SGD and Adam.

Interpretable Complex-Valued Neural Networks for Privacy Protection

- Liyao Xiang, Hao Zhang, Haotian Ma, Yifan Zhang, Jie Ren, Quanshi Zhang
- abstract@open-review(Poster): Previous studies have found that an adversary attacker can often infer unintended input information from intermediate-layer features. We study the possibility of preventing such adversarial inference, yet without too much accuracy degradation. We propose a generic method to revise the neural network to boost the challenge of inferring input attributes from features, while maintaining highly accurate outputs. In particular, the method transforms real-valued features into complex-valued ones, in which the input is hidden in a randomized phase of the transformed features. The knowledge of the phase acts like a key, with which any party can easily recover the output from the processing result, but without which the party can neither recover the output nor distinguish the original input. Preliminary experiments on various datasets and network structures have shown that our method significantly diminishes the adversary's ability in inferring about the input while largely preserves the resulting accuracy.

V-MPO: On-Policy Maximum a Posteriori Policy Optimization for Discrete and Continuous Control

- H. Francis Song, Abbas Abdolmaleki, Jost Tobias Springenberg, Aidan Clark, Hubert Soyer, Jack W. Rae, Seb Noury, Arun Ahuja, Siqi Liu, Dhruva Tirumala, Nicolas Heess, Dan Belov, Martin Riedmiller, Matthew M. Botvinick
- abstract@open-review(Poster): Some of the most successful applications of deep reinforcement learning to challenging domains in discrete and continuous control have used policy gradient methods in the on-policy setting. However, policy gradients can suffer from large variance that may limit performance, and in practice require carefully tuned entropy regularization to prevent policy collapse. As an alternative to policy gradient algorithms, we introduce V-MPO, an on-policy adaptation of Maximum a Posteriori Policy Optimization (MPO) that performs policy iteration based on a learned state-value function. We show that V-MPO surpasses previously reported scores for both the Atari-57 and DMLab-30 benchmark suites in the multi-task setting, and does so reliably without importance weighting, entropy regularization, or population-based tuning of hyperparameters. On individual DMLab and Atari levels, the proposed algorithm can achieve scores that are substantially higher than has previously been reported. V-MPO is also applicable to problems with high-dimensional, continuous action spaces, which we demonstrate in the context of learning to control simulated humanoids with 22 degrees of freedom from full state observations and 56 degrees of freedom from pixel observations, as well as example OpenAI Gym tasks where V-MPO achieves substantially higher asymptotic scores than previously reported.

Improving Adversarial Robustness Requires Revisiting Misclassified Examples

- Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, Quanquan Gu
- abstract@open-review(Poster): Deep neural networks (DNNs) are vulnerable to adversarial examples crafted by imperceptible perturbations. A range of defense techniques have been proposed to improve DNN robustness to adversarial examples, among which adversarial training has been demonstrated to be the most effective. Adversarial training is often formulated as a min-max optimization problem, with the inner maximization for generating adversarial examples. However, there exists a simple, yet easily overlooked fact that adversarial examples are only defined on correctly classified (natural) examples, but inevitably, some (natural) examples will be misclassified during training. In this paper, we investigate the distinctive influence of misclassified and correctly classified examples on the final robustness of adversarial training. Specifically, we find that misclassified examples indeed have a significant impact on the final robustness. More surprisingly, we find that different maximization techniques on misclassified examples may have a negligible influence on the final robustness, while different minimization techniques are crucial. Motivated by the above discovery, we propose a new defense algorithm called {\em Misclassification Aware adveRsarial Training} (MART), which explicitly differentiates the misclassified and correctly classified examples during the training. We also propose a semi-supervised extension of MART, which can leverage the unlabeled data to further improve the robustness. Experimental results show that MART and its variant could significantly improve the state-of-the-art adversarial robustness.

DivideMix: Learning with Noisy Labels as Semi-supervised Learning

- Junnan Li, Richard Socher, Steven C.H. Hoi
- abstract@open-review(Poster): Deep neural networks are known to be annotation-hungry. Numerous efforts have been devoted to reducing the annotation cost when learning with deep networks. Two prominent directions include learning with noisy labels and semi-supervised learning by exploiting unlabeled data. In this work, we propose DivideMix, a novel framework for learning with noisy labels by leveraging semi-supervised learning techniques. In particular, DivideMix models the per-sample loss distribution with a mixture model to dynamically divide the training data into a labeled set with clean samples and an unlabeled set with noisy samples, and trains the model on both the labeled and unlabeled data in a semi-supervised manner. To avoid confirmation bias, we simultaneously train two diverged networks where each network uses the dataset division from the other network. During the semi-supervised training phase, we improve the MixMatch strategy by performing label co-refinement and label co-guessing on labeled and unlabeled samples, respectively. Experiments on multiple benchmark datasets demonstrate substantial improvements over state-of-the-art methods. Code is available at https://github.com/LiJunnan1992/DivideMix .

Data-dependent Gaussian Prior Objective for Language Generation

- Zuchao Li, Rui Wang, Kehai Chen, Masso Utiyama, Eiichiro Sumita, Zhuosheng Zhang, Hai Zhao
- abstract@open-review(Talk): For typical sequence prediction problems such as language generation, maximum likelihood estimation (MLE) has commonly been adopted as it encourages the predicted sequence most consistent with the ground-truth sequence to have the highest probability of occurring. However, MLE focuses on once-to-all matching between the predicted sequence and gold-standard, consequently treating all incorrect predictions as being equally incorrect. We refer to this drawback as {\it negative diversity ignorance} in this paper. Treating all incorrect predictions as equal unfairly downplays the nuance of these sequences' detailed token-wise structure. To counteract this, we augment the MLE loss by introducing an extra Kullback--Leibler divergence term derived by comparing a data-dependent Gaussian prior and the detailed training prediction. The proposed data-dependent Gaussian prior objective (D2GPo) is defined over a prior topological order of tokens and is poles apart from the data-independent Gaussian prior (L2 regularization) commonly adopted in smoothing the training of MLE. Experimental results show that the proposed method makes effective use of a more detailed prior in the data and has improved performance in typical language generation tasks, including supervised and unsupervised machine translation, text summarization, storytelling, and image captioning.

Fooling Detection Alone is Not Enough: Adversarial Attack against Multiple Object Tracking

- Yunhan Jia, Yantao Lu, Junjie Shen, Qi Alfred Chen, Hao Chen, Zhenyu Zhong, Tao Wei
- abstract@open-review(Poster): Recent work in adversarial machine learning started to focus on the visual perception in autonomous driving and studied Adversarial Examples (AEs) for object detection models. However, in such visual perception pipeline the detected objects must also be tracked, in a process called Multiple Object Tracking (MOT), to build the moving trajectories of surrounding obstacles. Since MOT is designed to be robust against errors in object detection, it poses a general challenge to existing attack techniques that blindly target objection detection: we find that a success rate of over 98% is needed for them to actually affect the tracking results, a requirement that no existing attack technique can satisfy. In this paper, we are the first

to study adversarial machine learning attacks against the complete visual perception pipeline in autonomous driving, and discover a novel attack technique, tracker hijacking, that can effectively fool MOT using AEs on object detection. Using our technique, successful AEs on as few as one single frame can move an existing object in to or out of the headway of an autonomous vehicle to cause potential safety hazards. We perform evaluation using the Berkeley Deep Drive dataset and find that on average when 3 frames are attacked, our attack can have a nearly 100% success rate while attacks that blindly target object detection only have up to 25%.

Watch, Try, Learn: Meta-Learning from Demonstrations and Rewards

- Allan Zhou, Eric Jang, Daniel Kappler, Alex Herzog, Mohi Khansari, Paul Wohlhart, Yunfei Bai, Mrinal Kalakrishnan, Sergey Levine, Chelsea Finn
- abstract@open-review(Poster): Imitation learning allows agents to learn complex behaviors from demonstrations. However, learning a complex vision-based task may require an impractical number of demonstrations. Meta-imitation learning is a promising approach towards enabling agents to learn a new task from one or a few demonstrations by leveraging experience from learning similar tasks. In the presence of task ambiguity or unobserved dynamics, demonstrations alone may not provide enough information; an agent must also try the task to successfully infer a policy. In this work, we propose a method that can learn to learn from both demonstrations and trial-and-error experience with sparse reward feedback. In comparison to meta-imitation, this approach enables the agent to effectively and efficiently improve itself autonomously beyond the demonstration data. In comparison to meta-reinforcement learning, we can scale to substantially broader distributions of tasks, as the demonstration reduces the burden of exploration. Our experiments show that our method significantly outperforms prior approaches on a set of challenging, vision-based control tasks.

Logic and the 2-Simplicial Transformer

- James Clift, Dmitry Doryn, Daniel Murfet, James Wallbridge
- abstract@open-review(Poster): We introduce the 2-simplicial Transformer, an extension of the Transformer which includes a form of higher-dimensional attention generalising the dot-product attention, and uses this attention to update entity representations with tensor products of value vectors. We show that this architecture is a useful inductive bias for logical reasoning in the context of deep reinforcement learning.

AE-OT: A NEW GENERATIVE MODEL BASED ON EXTENDED SEMI-DISCRETE OPTIMAL TRANSPORT

- Dongsheng An, Yang Guo, Na Lei, Zhongxuan Luo, Shing-Tung Yau, Xianfeng Gu
- abstract@open-review(Poster): Generative adversarial networks (GANs) have attracted huge attention due to its capability to generate visual realistic images. However, most of the existing models suffer from the mode collapse or mode mixture problems. In this work, we give a theoretic explanation of the both problems by Figalli's regularity theory of optimal transportation maps. Basically, the generator compute the transportation maps between the white noise distributions and the data distributions, which are in general discontinuous. However, DNNs can only represent continuous maps. This intrinsic conflict induces mode collapse and mode mixture. In order to tackle the both problems, we explicitly separate the manifold embedding and the optimal transportation; the first part is carried out using an autoencoder to map the images onto the latent space; the second part is accomplished using a GPU-based convex optimization to find the discontinuous transportation maps. Composing the extended OT map and the decoder, we can finally generate new images from the white noise. This AE-OT model avoids representing discontinuous maps by DNNs, therefore effectively prevents mode collapse and mode mixture.

Enhancing Adversarial Defense by k-Winners-Take-All

- Chang Xiao, Peilin Zhong, Changxi Zheng
- abstract@open-review(Spotlight): We propose a simple change to existing neural network structures for better defending against gradient-based adversarial attacks. Instead of using popular activation functions (such as ReLU), we advocate the use of k-Winners-Take-All (k-WTA) activation, a C0 discontinuous function that purposely invalidates the neural network model's gradient at densely distributed input data points. The proposed k-WTA activation can be readily used in nearly all existing networks and training methods with no significant overhead. Our proposal is theoretically rationalized. We analyze why the discontinuities in k-WTA networks can largely prevent gradient-based search of adversarial examples and why they at the same time remain innocuous to the network training. This understanding is also empirically backed. We test k-WTA activation on various network structures optimized by a training method, be it adversarial training or not. In all cases, the robustness of k-WTA networks outperforms that of traditional networks under white-box attacks.

Exploration in Reinforcement Learning with Deep Covering Options

- Yuu Jinnai, Jee Won Park, Marlos C. Machado, George Konidaris
- abstract@open-review(Poster): While many option discovery methods have been proposed to accelerate exploration in reinforcement learning, they are often heuristic. Recently, covering options was proposed to discover a set of options that provably reduce the upper bound of the environment's cover time, a measure of the difficulty of exploration. Covering options are computed using the eigenvectors of the graph Laplacian, but they are constrained to tabular tasks and are not applicable to tasks with large or continuous state-spaces. We introduce deep covering options, an online method that extends covering options to large state spaces, automatically discovering task-agnostic options that encourage exploration. We evaluate our method in several challenging sparse-reward domains and we show that our approach identifies less explored regions of the state-space and successfully generates options to visit these regions, substantially improving both the exploration and the total accumulated reward.

Learning Disentangled Representations for CounterFactual Regression

- Negar Hassanpour, Russell Greiner
- abstract@open-review(Poster): We consider the challenge of estimating treatment effects from observational data; and point out that, in general, only some factors based on the observed covariates X contribute to selection of the treatment T, and only some to determining the outcomes Y. We model this by considering three underlying sources of {X, T, Y} and show that explicitly modeling these sources offers great insight to guide designing models that better handle selection bias. This paper is an attempt to conceptualize this line of thought and provide a path to explore it further. In this work, we propose an algorithm to (1) identify disentangled representations of the above-mentioned underlying factors from any given observational dataset D and (2) leverage this knowledge to reduce, as well as account for, the negative impact of selection bias on estimating the treatment effects from D. Our empirical results show that the proposed method achieves state-of-the-art performance in both individual and population based evaluation measures.

Analysis of Video Feature Learning in Two-Stream CNNs on the Example of Zebrafish Swim Bout Classification

- Bennet Breier, Arno Onken
- abstract@open-review(Poster): Semmelhack et al. (2014) have achieved high classification accuracy in distinguishing swim bouts of zebrafish using a Support Vector Machine (SVM). Convolutional Neural Networks (CNNs) have reached superior performance in various image recognition tasks over SVMs, but these powerful networks remain a black box. Reaching better transparency helps to build trust in their classifications and makes learned features interpretable to experts. Using a recently developed technique called Deep Taylor Decomposition, we generated heatmaps to highlight input regions of high relevance for predictions. We find that our CNN makes predictions by analyzing the steadiness of the tail's trunk, which markedly differs from the manually extracted features used by Semmelhack et al. (2014). We further uncovered that the network paid attention to experimental artifacts.

Removing these artifacts ensured the validity of predictions. After correction, our best CNN beats the SVM by 6.12%, achieving a classification accuracy of 96.32%. Our work thus demonstrates the utility of AI explainability for CNNs.

Robust Local Features for Improving the Generalization of Adversarial Training

- Chuanbiao Song, Kun He, Jiadong Lin, Liwei Wang, John E. Hopcroft
- abstract@open-review(Poster): Adversarial training has been demonstrated as one of the most effective methods for training robust models to defend against adversarial examples. However, adversarially trained models often lack adversarially robust generalization on unseen testing data. Recent works show that adversarially trained models are more biased towards global structure features. Instead, in this work, we would like to investigate the relationship between the generalization of adversarial training and the robust local features, as the robust local features generalize well for unseen shape variation. To learn the robust local features, we develop a Random Block Shuffle (RBS) transformation to break up the global structure features on normal adversarial examples. We continue to propose a new approach called Robust Local Features for Adversarial Training (RLFAT), which first learns the robust local features by adversarial training on the RBS-transformed adversarial examples, and then transfers the robust local features into the training of normal adversarial examples. To demonstrate the generality of our argument, we implement RLFAT in currently state-of-the-art adversarial training frameworks. Extensive experiments on STL-10, CIFAR-10 and CIFAR-100 show that RLFAT significantly improves both the adversarially robust generalization and the standard generalization of adversarial training. Additionally, we demonstrate that our models capture more local features of the object on the images, aligning better with human perception.

Online and stochastic optimization beyond Lipschitz continuity: A Riemannian approach

- Kimon Antonakopoulos, E. Veronica Belmega, Panayotis Mertikopoulos
- abstract@open-review(Spotlight): Motivated by applications to machine learning and imaging science, we study a class of online and stochastic optimization problems with loss functions that are not Lipschitz continuous; in particular, the loss functions encountered by the optimizer could exhibit gradient singularities or be singular themselves. Drawing on tools and techniques from Riemannian geometry, we examine a Riemann–Lipschitz (RL) continuity condition which is tailored to the singularity landscape of the problem's loss functions. In this way, we are able to tackle cases beyond the Lipschitz framework provided by a global norm, and we derive optimal regret bounds and last iterate convergence results through the use of regularized learning methods (such as online mirror descent). These results are subsequently validated in a class of stochastic Poisson inverse problems that arise in imaging science.

Reinforcement Learning with Competitive Ensembles of Information-Constrained Primitives

- Anirudh Goyal, Shagun Sodhani, Jonathan Binas, Xue Bin Peng, Sergey Levine, Yoshua Bengio
- abstract@open-review(Poster): Reinforcement learning agents that operate in diverse and complex environments can benefit from the structured decomposition of their behavior. Often, this is addressed in the context of hierarchical reinforcement learning, where the aim is to decompose a policy into lower-level primitives or options, and a higher-level meta-policy that triggers the appropriate behaviors for a given situation. However, the meta-policy must still produce appropriate decisions in all states. In this work, we propose a policy design that decomposes into primitives, similarly to hierarchical reinforcement learning, but without a high-level meta-policy. Instead, each primitive can decide for themselves whether they wish to act in the current state. We use an information-theoretic mechanism for enabling this decentralized decision: each primitive chooses how much information it needs about the current state to make a decision and the primitive that requests the most information about the current state acts in the world. The primitives are regularized to use as little information as possible, which leads to natural competition and specialization. We experimentally demonstrate that this policy architecture improves over both flat and hierarchical policies in terms of generalization.

Learning the Arrow of Time for Problems in Reinforcement Learning

- Nasim Rahaman, Steffen Wolf, Anirudh Goyal, Roman Remme, Yoshua Bengio
- abstract@open-review(Poster): We humans have an innate understanding of the asymmetric progression of time, which we use to efficiently and safely perceive and manipulate our environment. Drawing inspiration from that, we approach the problem of learning an arrow of time in a Markov (Decision) Process. We illustrate how a learned arrow of time can capture salient information about the environment, which in turn can be used to measure reachability, detect side-effects and to obtain an intrinsic reward signal. Finally, we propose a simple yet effective algorithm to parameterize the problem at hand and learn an arrow of time with a function approximator (here, a deep neural network). Our empirical results span a selection of discrete and continuous environments, and demonstrate for a class of stochastic processes that the learned arrow of time agrees reasonably well with a well known notion of an arrow of time due to Jordan, Kinderlehrer and Otto (1998).

The Variational Bandwidth Bottleneck: Stochastic Evaluation on an Information Budget

- Anirudh Goyal, Yoshua Bengio, Matthew Botvinick, Sergey Levine
- abstract@open-review(Poster): In many applications, it is desirable to extract only the relevant information from complex input data, which involves making a decision about which input features are relevant. The information bottleneck method formalizes this as an information-theoretic optimization problem by maintaining an optimal tradeoff between compression (throwing away irrelevant input information), and predicting the target. In many problem settings, including the reinforcement learning problems we consider in this work, we might prefer to compress only part of the input. This is typically the case when we have a standard conditioning input, such as a state observation, and a ``privileged'' input, which might correspond to the goal of a task, the output of a costly planning algorithm, or communication with another agent. In such cases, we might prefer to compress the privileged input, either to achieve better generalization (e.g., with respect to goals) or to minimize access to costly information (e.g., in the case of communication). Practical implementations of the information bottleneck based on variational inference require access to the privileged input in order to compute the bottleneck variable, so although they perform compression, this compression operation itself needs unrestricted, lossless access. In this work, we propose the variational bandwidth bottleneck, which decides for each example on the estimated value of the privileged information before seeing it, i.e., only based on the standard input, and then accordingly chooses stochastically, whether to access the privileged input or not. We formulate a tractable approximation to this framework and demonstrate in a series of reinforcement learning experiments that it can improve generalization and reduce access to computationally costly information.

The Implicit Bias of Depth: How Incremental Learning Drives Generalization

- Daniel Gissin, Shai Shalev-Shwartz, Amit Daniely
- abstract@open-review(Poster): A leading hypothesis for the surprising generalization of neural networks is that the dynamics of gradient descent bias the model towards simple solutions, by searching through the solution space in an incremental order of complexity. We formally define the notion of incremental learning dynamics and derive the conditions on depth and initialization for which this phenomenon arises in deep linear models. Our main theoretical contribution is a dynamical depth separation result, proving that while shallow models can exhibit incremental learning dynamics, they require the initialization to be exponentially small for these dynamics to present themselves. However, once the model becomes deeper, the dependence becomes polynomial and incremental learning can arise in more natural settings. We complement our theoretical findings by experimenting with deep matrix sensing, quadratic neural networks and with binary classification using diagonal and convolutional linear networks, showing all of these models exhibit incremental learning.

Rethinking Softmax Cross-Entropy Loss for Adversarial Robustness

- Tianyu Pang, Kun Xu, Yinpeng Dong, Chao Du, Ning Chen, Jun Zhu
- abstract@open-review(Poster): Previous work shows that adversarially robust generalization requires larger sample complexity, and the same dataset, e.g., CIFAR-10, which enables good standard accuracy may not suffice to train robust models. Since collecting new training data could be costly, we focus on better utilizing the given data by inducing the regions with high sample density in the feature space, which could lead to locally sufficient samples for robust learning. We first formally show that the softmax cross-entropy (SCE) loss and its variants convey inappropriate supervisory signals, which encourage the learned feature points to spread over the space sparsely in training. This inspires us to propose the Max-Mahalanobis center (MMC) loss to explicitly induce dense feature regions in order to benefit robustness. Namely, the MMC loss encourages the model to concentrate on learning ordered and compact representations, which gather around the preset optimal centers for different classes. We empirically demonstrate that applying the MMC loss can significantly improve robustness even under strong adaptive attacks, while keeping state-of-the-art accuracy on clean inputs with little extra computation compared to the SCE loss.

Measuring Compositional Generalization: A Comprehensive Method on Realistic Data

- Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, Olivier Bousquet
- abstract@open-review(Poster): State-of-the-art machine learning methods exhibit limited compositional generalization. At the same time, there is a lack of realistic benchmarks that comprehensively measure this ability, which makes it challenging to find and evaluate improvements. We introduce a novel method to systematically construct such benchmarks by maximizing compound divergence while guaranteeing a small atom divergence between train and test sets, and we quantitatively compare this method to other approaches for creating compositional generalization benchmarks. We present a large and realistic natural language question answering dataset that is constructed according to this method, and we use it to analyze the compositional generalization ability of three machine learning architectures. We find that they fail to generalize compositionally and that there is a surprisingly strong negative correlation between compound divergence and accuracy. We also demonstrate how our method can be used to create new compositionality benchmarks on top of the existing SCAN dataset, which confirms these findings.

Theory and Evaluation Metrics for Learning Disentangled Representations

- Kien Do, Truyen Tran
- abstract@open-review(Poster): We make two theoretical contributions to disentanglement learning by (a) defining precise semantics of disentangled representations, and (b) establishing robust metrics for evaluation. First, we characterize the concept “disentangled representations” used in supervised and unsupervised methods along three dimensions–informativeness, separability and interpretability–which can be expressed and quantified explicitly using information-theoretic constructs. This helps explain the behaviors of several well-known disentanglement learning models. We then propose robust metrics for measuring informativeness, separability and interpretability. Through a comprehensive suite of experiments, we show that our metrics correctly characterize the representations learned by different methods and are consistent with qualitative (visual) results. Thus, the metrics allow disentanglement learning methods to be compared on a fair ground. We also empirically uncovered new interesting properties of VAE-based methods and interpreted them with our formulation. These findings are promising and hopefully will encourage the design of more theoretically driven models for learning disentangled representations.

Mixup Inference: Better Exploiting Mixup to Defend Adversarial Attacks

- Tianyu Pang, *Kun Xu*, Jun Zhu
- abstract@open-review(Poster): It has been widely recognized that adversarial examples can be easily crafted to fool deep networks, which mainly root from the locally non-linear behavior nearby input examples. Applying mixup in training provides an effective mechanism to improve generalization performance and model robustness against adversarial perturbations, which introduces the globally linear behavior in-between training examples. However, in previous work, the mixup-trained models only passively defend adversarial attacks in inference by directly classifying the inputs, where the induced global linearity is not well exploited. Namely, since the locality of the adversarial perturbations, it would be more efficient to actively break the locality via the globality of the model predictions. Inspired by simple geometric intuition, we develop an inference principle, named mixup inference (MI), for mixup-trained models. MI mixups the input with other random clean samples, which can shrink and transfer the equivalent perturbation if the input is adversarial. Our experiments on CIFAR-10 and CIFAR-100 demonstrate that MI can further improve the adversarial robustness for the models trained by mixup and its variants.

Dynamically Pruned Message Passing Networks for Large-scale Knowledge Graph Reasoning

- Xiaoran Xu, Wei Feng, Yunsheng Jiang, Xiaohui Xie, Zhiqing Sun, Zhi-Hong Deng
- abstract@open-review(Poster): We propose Dynamically Pruned Message Passing Networks (DPMPN) for large-scale knowledge graph reasoning. In contrast to existing models, embedding-based or path-based, we learn an input-dependent subgraph to explicitly model a sequential reasoning process. Each subgraph is dynamically constructed, expanding itself selectively under a flow-style attention mechanism. In this way, we can not only construct graphical explanations to interpret prediction, but also prune message passing in Graph Neural Networks (GNNs) to scale with the size of graphs. We take the inspiration from the consciousness prior proposed by Bengio to design a two-GNN framework to encode global input-invariant graph-structured representation and learn local input-dependent one coordinated by an attention module. Experiments show the reasoning capability in our model that is providing a clear graphical explanation as well as predicting results accurately, outperforming most state-of-the-art methods in knowledge base completion tasks.

Are Pre-trained Language Models Aware of Phrases? Simple but Strong Baselines for Grammar Induction

- Taeuk Kim, Jihun Choi, Daniel Edmiston, Sang-goo Lee
- abstract@open-review(Poster): With the recent success and popularity of pre-trained language models (LMs) in natural language processing, there has been a rise in efforts to understand their inner workings. In line with such interest, we propose a novel method that assists us in investigating the extent to which pre-trained LMs capture the syntactic notion of constituency. Our method provides an effective way of extracting constituency trees from the pre-trained LMs without training. In addition, we report intriguing findings in the induced trees, including the fact that pre-trained LMs outperform other approaches in correctly demarcating adverb phrases in sentences.

FSPool: Learning Set Representations with Featurewise Sort Pooling

- Yan Zhang, Jonathon Hare, Adam Prügel-Bennett
- abstract@open-review(Poster): Traditional set prediction models can struggle with simple datasets due to an issue we call the responsibility problem. We introduce a pooling method for sets of feature vectors based on sorting features across elements of the set. This can be used to construct a permutation-equivariant auto-encoder that avoids this responsibility problem. On a toy dataset of polygons and a set version of MNIST, we show that such an auto-encoder produces considerably better reconstructions and representations. Replacing the pooling function in existing set encoders with FSPool improves accuracy and convergence speed on a variety of datasets.

On the Convergence of FedAvg on Non-IID Data

- Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, Zhihua Zhang
- abstract@open-review(Talk): Federated learning enables a large amount of edge computing devices to jointly learn a model without data sharing. As a leading algorithm in this setting, Federated Averaging (\texttt{FedAvg}) runs Stochastic Gradient Descent (SGD) in parallel on a small subset of the total devices and averages the sequences only once in a while. Despite its simplicity, it lacks theoretical guarantees under realistic settings. In this paper, we analyze the convergence of \texttt{FedAvg} on non-iid data and establish a convergence rate of $\mathcal{O}(\frac{1}{T})$ for strongly convex and smooth problems, where T is the number of SGDs. Importantly, our bound demonstrates a trade-off between communication-efficiency and convergence rate. As user devices may be disconnected from the server, we relax the assumption of full device participation to partial device participation and study different averaging schemes; low device participation rate can be achieved without severely slowing down the learning. Our results indicate that heterogeneity of data slows down the convergence, which matches empirical observations. Furthermore, we provide a necessary condition for \texttt{FedAvg} on non-iid data: the learning rate η must decay, even if full-gradient is used; otherwise, the solution will be $\Omega (\eta)$ away from the optimal.

Multi-agent Reinforcement Learning for Networked System Control

- Tianshu Chu, Sandeep Chinchali, Sachin Katti
- abstract@open-review(Poster): This paper considers multi-agent reinforcement learning (MARL) in networked system control. Specifically, each agent learns a decentralized control policy based on local observations and messages from connected neighbors. We formulate such a networked MARL (NMARL) problem as a spatiotemporal Markov decision process and introduce a spatial discount factor to stabilize the training of each local agent. Further, we propose a new differentiable communication protocol, called NeurComm, to reduce information loss and non-stationarity in NMARL. Based on experiments in realistic NMARL scenarios of adaptive traffic signal control and cooperative adaptive cruise control, an appropriate spatial discount factor effectively enhances the learning curves of non-communicative MARL algorithms, while NeurComm outperforms existing communication protocols in both learning efficiency and control performance.

Hierarchical Foresight: Self-Supervised Learning of Long-Horizon Tasks via Visual Subgoal Generation

- Suraj Nair, Chelsea Finn
- abstract@open-review(Poster): Video prediction models combined with planning algorithms have shown promise in enabling robots to learn to perform many vision-based tasks through only self-supervision, reaching novel goals in cluttered scenes with unseen objects. However, due to the compounding uncertainty in long horizon video prediction and poor scalability of sampling-based planning optimizers, one significant limitation of these approaches is the ability to plan over long horizons to reach distant goals. To that end, we propose a framework for subgoal generation and planning, hierarchical visual foresight (HVF), which generates subgoal images conditioned on a goal image, and uses them for planning. The subgoal images are directly optimized to decompose the task into easy to plan segments, and as a result, we observe that the method naturally identifies semantically meaningful states as subgoals. Across three out of four simulated vision-based manipulation tasks, we find that our method achieves more than 20% absolute performance improvement over planning without subgoals and model-free RL approaches. Further, our experiments illustrate that our approach extends to real, cluttered visual scenes.

Neural Stored-program Memory

- Hung Le, Truyen Tran, Svetha Venkatesh
- abstract@open-review(Poster): Neural networks powered with external memory simulate computer behaviors. These models, which use the memory to store data for a neural controller, can learn algorithms and other complex tasks. In this paper, we introduce a new memory to store weights for the controller, analogous to the stored-program memory in modern computer architectures. The proposed model, dubbed Neural Stored-program Memory, augments current memory-augmented neural networks, creating differentiable machines that can switch programs through time, adapt to variable contexts and thus fully resemble the Universal Turing Machine. A wide range of experiments demonstrate that the resulting machines not only excel in classical algorithmic problems, but also have potential for compositional, continual, few-shot learning and question-answering tasks.

ES-MAML: Simple Hessian-Free Meta Learning

- Xingyou Song, Wenbo Gao, Yuxiang Yang, Krzysztof Choromanski, Aldo Pacchiano, Yunhao Tang
- abstract@open-review(Poster): We introduce ES-MAML, a new framework for solving the model agnostic meta learning (MAML) problem based on Evolution Strategies (ES). Existing algorithms for MAML are based on policy gradients, and incur significant difficulties when attempting to estimate second derivatives using backpropagation on stochastic policies. We show how ES can be applied to MAML to obtain an algorithm which avoids the problem of estimating second derivatives, and is also conceptually simple and easy to implement. Moreover, ES-MAML can handle new types of nonsmooth adaptation operators, and other techniques for improving performance and estimation of ES methods become applicable. We show empirically that ES-MAML is competitive with existing methods and often yields better adaptation with fewer queries.

TabFact: A Large-scale Dataset for Table-based Fact Verification

- Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, William Yang Wang
- abstract@open-review(Poster): The problem of verifying whether a textual hypothesis holds based on the given evidence, also known as fact verification, plays an important role in the study of natural language understanding and semantic representation. However, existing studies are mainly restricted to dealing with unstructured evidence (e.g., natural language sentences and documents, news, etc), while verification under structured evidence, such as tables, graphs, and databases, remains unexplored. This paper specifically aims to study the fact verification given semi-structured data as evidence. To this end, we construct a large-scale dataset called TabFact with 16k Wikipedia tables as the evidence for 118k human-annotated natural language statements, which are labeled as either ENTAILED or REFUTED. TabFact is challenging since it involves both soft linguistic reasoning and hard symbolic reasoning. To address these reasoning challenges, we design two different models: Table-BERT and Latent Program Algorithm (LPA). Table-BERT leverages the state-of-the-art pre-trained language model to encode the linearized tables and statements into continuous vectors for verification. LPA parses statements into LISP-like programs and executes them against the tables to obtain the returned binary value for verification. Both methods achieve similar accuracy but still lag far behind human performance. We also perform a comprehensive analysis to demonstrate great future opportunities.

Implicit Bias of Gradient Descent based Adversarial Training on Separable Data

- Yan Li, Ethan X.Fang, Huan Xu, Tuo Zhao
- abstract@open-review(Poster): Adversarial training is a principled approach for training robust neural networks. Despite of tremendous successes in practice, its theoretical properties still remain largely unexplored. In this paper, we provide new theoretical insights of gradient descent based adversarial training by studying its computational properties, specifically on its implicit bias. We take the binary classification task on linearly separable data as an illustrative example, where the loss asymptotically attains its infimum as the parameter diverges to infinity along certain directions. Specifically, we show that for any fixed iteration T, when the adversarial perturbation during training has proper bounded L2 norm, the classifier learned by gradient descent based adversarial training converges in direction to the maximum L2 norm margin classifier at the rate of $O(1/\sqrt{T})$, significantly faster than the rate

$O(1/\log T)$ of training with clean data. In addition, when the adversarial perturbation during training has bounded Lq norm, the resulting classifier converges in direction to a maximum mixed-norm margin classifier, which has a natural interpretation of robustness, as being the maximum L2 norm margin classifier under worst-case bounded Lq norm perturbation to the data. Our findings provide theoretical backups for adversarial training that it indeed promotes robustness against adversarial perturbation.

Image-guided Neural Object Rendering

- Justus Thies, Michael Zollhöfer, Christian Theobalt, Marc Stamminger, Matthias Nießner
- abstract@open-review(Poster): We propose a learned image-guided rendering technique that combines the benefits of image-based rendering and GAN-based image synthesis. The goal of our method is to generate photo-realistic re-renderings of reconstructed objects for virtual and augmented reality applications (e.g., virtual showrooms, virtual tours and sightseeing, the digital inspection of historical artifacts). A core component of our work is the handling of view-dependent effects. Specifically, we directly train an object-specific deep neural network to synthesize the view-dependent appearance of an object. As input data we are using an RGB video of the object. This video is used to reconstruct a proxy geometry of the object via multi-view stereo. Based on this 3D proxy, the appearance of a captured view can be warped into a new target view as in classical image-based rendering. This warping assumes diffuse surfaces, in case of view-dependent effects, such as specular highlights, it leads to artifacts. To this end, we propose EffectsNet, a deep neural network that predicts view-dependent effects. Based on these estimations, we are able to convert observed images to diffuse images. These diffuse images can be projected into other views. In the target view, our pipeline reinserts the new view-dependent effects. To composite multiple reprojected images to a final output, we learn a composition network that outputs photo-realistic results. Using this image-guided approach, the network does not have to allocate capacity on ``remembering" object appearance, instead it learns how to combine the appearance of captured images. We demonstrate the effectiveness of our approach both qualitatively and quantitatively on synthetic as well as on real data.

PC-DARTS: Partial Channel Connections for Memory-Efficient Architecture Search

- Yuhui Xu, Lingxi Xie, Xiaopeng Zhang, Xin Chen, Guo-Jun Qi, Qi Tian, Hongkai Xiong
- abstract@open-review(Spotlight): Differentiable architecture search (DARTS) provided a fast solution in finding effective network architectures, but suffered from large memory and computing overheads in jointly training a super-net and searching for an optimal architecture. In this paper, we present a novel approach, namely Partially-Connected DARTS, by sampling a small part of super-net to reduce the redundancy in exploring the network space, thereby performing a more efficient search without comprising the performance. In particular, we perform operation search in a subset of channels while bypassing the held out part in a shortcut. This strategy may suffer from an undesired inconsistency on selecting the edges of super-net caused by sampling different channels. We solve it by introducing edge normalization, which adds a new set of edge-level hyper-parameters to reduce uncertainty in search. Thanks to the reduced memory cost, PC-DARTS can be trained with a larger batch size and, consequently, enjoy both faster speed and higher training stability. Experiment results demonstrate the effectiveness of the proposed method. Specifically, we achieve an error rate of 2.57% on CIFAR10 within merely 0.1 GPU-days for architecture search, and a state-of-the-art top-1 error rate of 24.2% on ImageNet (under the mobile setting) within 3.8 GPU-days for search. Our code has been made available at https://www.dropbox.com/sh/on9lg3rpx1r6dkf/AABG5mt0sMHjnEJyoRnLEYW4a?dl=0.

Knowledge Consistency between Neural Networks and Beyond

- Ruofan Liang, Tianlin Li, Longfei Li, Jing Wang, Quanshi Zhang
- abstract@open-review(Poster): This paper aims to analyze knowledge consistency between pre-trained deep neural networks. We propose a generic definition for knowledge consistency between neural networks at different fuzziness levels. A task-agnostic method is designed to disentangle feature components, which represent the consistent knowledge, from raw intermediate-layer features of each neural network. As a generic tool, our method can be broadly used for different applications. In preliminary experiments, we have used knowledge consistency as a tool to diagnose representations of neural networks. Knowledge consistency provides new insights to explain the success of existing deep-learning techniques, such as knowledge distillation and network compression. More crucially, knowledge consistency can also be used to refine pre-trained networks and boost performance.

Lazy-CFR: fast and near-optimal regret minimization for extensive games with imperfect information

- Yichi Zhou, Tongzheng Ren, Jialian Li, Dong Yan, Jun Zhu
- abstract@open-review(Poster): Counterfactual regret minimization (CFR) methods are effective for solving two-player zero-sum extensive games with imperfect information with state-of-the-art results. However, the vanilla CFR has to traverse the whole game tree in each round, which is time-consuming in large-scale games. In this paper, we present Lazy-CFR, a CFR algorithm that adopts a lazy update strategy to avoid traversing the whole game tree in each round. We prove that the regret of Lazy-CFR is almost the same to the regret of the vanilla CFR and only needs to visit a small portion of the game tree. Thus, Lazy-CFR is provably faster than CFR. Empirical results consistently show that Lazy-CFR is significantly faster than the vanilla CFR.

Principled Weight Initialization for Hypernetworks

- Oscar Chang, Lampros Flokas, Hod Lipson
- abstract@open-review(Talk): Hypernetworks are meta neural networks that generate weights for a main neural network in an end-to-end differentiable manner. Despite extensive applications ranging from multi-task learning to Bayesian deep learning, the problem of optimizing hypernetworks has not been studied to date. We observe that classical weight initialization methods like Glorot & Bengio (2010) and He et al. (2015), when applied directly on a hypernet, fail to produce weights for the mainnet in the correct scale. We develop principled techniques for weight initialization in hypernets, and show that they lead to more stable mainnet weights, lower training loss, and faster convergence.

Additive Powers-of-Two Quantization: An Efficient Non-uniform Discretization for Neural Networks

- Yuhang Li, Xin Dong, Wei Wang
- abstract@open-review(Poster): We propose Additive Powers-of-Two~(APoT) quantization, an efficient non-uniform quantization scheme for the bell-shaped and long-tailed distribution of weights and activations in neural networks. By constraining all quantization levels as the sum of Powers-of-Two terms, APoT quantization enjoys high computational efficiency and a good match with the distribution of weights. A simple reparameterization of the clipping function is applied to generate a better-defined gradient for learning the clipping threshold. Moreover, weight normalization is presented to refine the distribution of weights to make the training more stable and consistent. Experimental results show that our proposed method outperforms state-of-the-art methods, and is even competitive with the full-precision models, demonstrating the effectiveness of our proposed APoT quantization. For example, our 4-bit quantized ResNet-50 on ImageNet achieves 76.6% top-1 accuracy without bells and whistles; meanwhile, our model reduces 22% computational cost compared with the uniformly quantized counterpart.

Enhancing Transformation-Based Defenses Against Adversarial Attacks with a Distribution Classifier

- Connie Kou, Hwee Kuan Lee, Ee-Chien Chang, Teck Khim Ng
- abstract@open-review(Poster): Adversarial attacks on convolutional neural networks (CNN) have gained significant attention and there have been active research efforts on defense mechanisms. Stochastic input transformation methods have been proposed, where the idea is to recover the image from adversarial attack by random transformation, and to take the majority vote as consensus among the random samples. However, the transformation

improves the accuracy on adversarial images at the expense of the accuracy on clean images. While it is intuitive that the accuracy on clean images would deteriorate, the exact mechanism in which how this occurs is unclear. In this paper, we study the distribution of softmax induced by stochastic transformations. We observe that with random transformations on the clean images, although the mass of the softmax distribution could shift to the wrong class, the resulting distribution of softmax could be used to correct the prediction. Furthermore, on the adversarial counterparts, with the image transformation, the resulting shapes of the distribution of softmax are similar to the distributions from the clean images. With these observations, we propose a method to improve existing transformation-based defenses. We train a separate lightweight distribution classifier to recognize distinct features in the distributions of softmax outputs of transformed images. Our empirical studies show that our distribution classifier, by training on distributions obtained from clean images only, outperforms majority voting for both clean and adversarial images. Our method is generic and can be integrated with existing transformation-based defenses.

Observational Overfitting in Reinforcement Learning

- Xingyou Song, Yiding Jiang, Stephen Tu, Yilun Du, Behnam Neyshabur
- abstract@open-review(Poster): A major component of overfitting in model-free reinforcement learning (RL) involves the case where the agent may mistakenly correlate reward with certain spurious features from the observations generated by the Markov Decision Process (MDP). We provide a general framework for analyzing this scenario, which we use to design multiple synthetic benchmarks from only modifying the observation space of an MDP. When an agent overfits to different observation spaces even if the underlying MDP dynamics is fixed, we term this observational overfitting. Our experiments expose intriguing properties especially with regards to implicit regularization, and also corroborate results from previous works in RL generalization and supervised learning (SL).

On Mutual Information Maximization for Representation Learning

- Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, Mario Lucic
- abstract@open-review(Poster): Many recent methods for unsupervised or self-supervised representation learning train feature extractors by maximizing an estimate of the mutual information (MI) between different views of the data. This comes with several immediate problems: For example, MI is notoriously hard to estimate, and using it as an objective for representation learning may lead to highly entangled representations due to its invariance under arbitrary invertible transformations. Nevertheless, these methods have been repeatedly shown to excel in practice. In this paper we argue, and provide empirical evidence, that the success of these methods cannot be attributed to the properties of MI alone, and that they strongly depend on the inductive bias in both the choice of feature extractor architectures and the parametrization of the employed MI estimators. Finally, we establish a connection to deep metric learning and argue that this interpretation may be a plausible explanation for the success of the recently introduced methods.

Tranquil Clouds: Neural Networks for Learning Temporally Coherent Features in Point Clouds

- Lukas Prantl, Nuttapong Chentanez, Stefan Jeschke, Nils Thuerey
- abstract@open-review(Spotlight): Point clouds, as a form of Lagrangian representation, allow for powerful and flexible applications in a large number of computational disciplines. We propose a novel deep-learning method to learn stable and temporally coherent feature spaces for points clouds that change over time. We identify a set of inherent problems with these approaches: without knowledge of the time dimension, the inferred solutions can exhibit strong flickering, and easy solutions to suppress this flickering can result in undesirable local minima that manifest themselves as halo structures. We propose a novel temporal loss function that takes into account higher time derivatives of the point positions, and encourages mingling, i.e., to prevent the aforementioned halos. We combine these techniques in a super-resolution method with a truncation approach to flexibly adapt the size of the generated positions. We show that our method works for large, deforming point sets from different sources to demonstrate the flexibility of our approach.

Ranking Policy Gradient

- Kaixiang Lin, Jiayu Zhou
- abstract@open-review(Poster): Sample inefficiency is a long-lasting problem in reinforcement learning (RL). The state-of-the-art estimates the optimal action values while it usually involves an extensive search over the state-action space and unstable optimization. Towards the sample-efficient RL, we propose ranking policy gradient (RPG), a policy gradient method that learns the optimal rank of a set of discrete actions. To accelerate the learning of policy gradient methods, we establish the equivalence between maximizing the lower bound of return and imitating a near-optimal policy without accessing any oracles. These results lead to a general off-policy learning framework, which preserves the optimality, reduces variance, and improves the sample-efficiency. We conduct extensive experiments showing that when consolidating with the off-policy learning framework, RPG substantially reduces the sample complexity, comparing to the state-of-the-art.

SVQN: Sequential Variational Soft Q-Learning Networks

- Shiyu Huang, Hang Su, Jun Zhu, Ting Chen
- abstract@open-review(Poster): Partially Observable Markov Decision Processes (POMDPs) are popular and flexible models for real-world decision-making applications that demand the information from past observations to make optimal decisions. Standard reinforcement learning algorithms for solving Markov Decision Processes (MDP) tasks are not applicable, as they cannot infer the unobserved states. In this paper, we propose a novel algorithm for POMDPs, named sequential variational soft Q-learning networks (SVQNs), which formalizes the inference of hidden states and maximum entropy reinforcement learning (MERL) under a unified graphical model and optimizes the two modules jointly. We further design a deep recurrent neural network to reduce the computational complexity of the algorithm. Experimental results show that SVQNs can utilize past information to help decision making for efficient inference, and outperforms other baselines on several challenging tasks. Our ablation study shows that SVQNs have the generalization ability over time and are robust to the disturbance of the observation.

Neural Machine Translation with Universal Visual Representation

- Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, Hai Zhao
- abstract@open-review(Spotlight): Though visual information has been introduced for enhancing neural machine translation (NMT), its effectiveness strongly relies on the availability of large amounts of bilingual parallel sentence pairs with manual image annotations. In this paper, we present a universal visual representation learned over the monolingual corpora with image annotations, which overcomes the lack of large-scale bilingual sentence-image pairs, thereby extending image applicability in NMT. In detail, a group of images with similar topics to the source sentence will be retrieved from a light topic-image lookup table learned over the existing sentence-image pairs, and then is encoded as image representations by a pre-trained ResNet. An attention layer with a gated weighting is to fuse the visual information and text information as input to the decoder for predicting target translations. In particular, the proposed method enables the visual information to be integrated into large-scale text-only NMT in addition to the multimodel NMT. Experiments on four widely used translation datasets, including the WMT'16 English-to-Romanian, WMT'14 English-to-German, WMT'14 English-to-French, and Multi30K, show that the proposed approach achieves significant improvements over strong baselines.

Understanding Architectures Learnt by Cell-based Neural Architecture Search

- Yao Shu, Wei Wang, Shaofeng Cai

- abstract@open-review(Poster): Neural architecture search (NAS) searches architectures automatically for given tasks, e.g., image classification and language modeling. Improving the search efficiency and effectiveness has attracted increasing attention in recent years. However, few efforts have been devoted to understanding the generated architectures. In this paper, we first reveal that existing NAS algorithms (e.g., DARTS, ENAS) tend to favor architectures with wide and shallow cell structures. These favorable architectures consistently achieve fast convergence and are consequently selected by NAS algorithms. Our empirical and theoretical study further confirms that their fast convergence derives from their smooth loss landscape and accurate gradient information. Nonetheless, these architectures may not necessarily lead to better generalization performance compared with other candidate architectures in the same search space, and therefore further improvement is possible by revising existing NAS algorithms.

AutoQ: Automated Kernel-Wise Neural Network Quantization

- Qian Lou, Feng Guo, Minje Kim, Lantao Liu, Lei Jiang.
- abstract@open-review(Poster): Network quantization is one of the most hardware friendly techniques to enable the deployment of convolutional neural networks (CNNs) on low-power mobile devices. Recent network quantization techniques quantize each weight kernel in a convolutional layer independently for higher inference accuracy, since the weight kernels in a layer exhibit different variances and hence have different amounts of redundancy. The quantization bitwidth or bit number (QBN) directly decides the inference accuracy, latency, energy and hardware overhead. To effectively reduce the redundancy and accelerate CNN inferences, various weight kernels should be quantized with different QBNs. However, prior works use only one QBN to quantize each convolutional layer or the entire CNN, because the design space of searching a QBN for each weight kernel is too large. The hand-crafted heuristic of the kernel-wise QBN search is so sophisticated that domain experts can obtain only sub-optimal results. It is difficult for even deep reinforcement learning (DRL) DDPG-based agents to find a kernel-wise QBN configuration that can achieve reasonable inference accuracy. In this paper, we propose a hierarchical-DRL-based kernel-wise network quantization technique, AutoQ, to automatically search a QBN for each weight kernel, and choose another QBN for each activation layer. Compared to the models quantized by the state-of-the-art DRL-based schemes, on average, the same models quantized by AutoQ reduce the inference latency by 54.06%, and decrease the inference energy consumption by 50.69%, while achieving the same inference accuracy.

White Noise Analysis of Neural Networks

- Ali Borji, Sikun Lin
- abstract@open-review(Spotlight): A white noise analysis of modern deep neural networks is presented to unveil their biases at the whole network level or the single neuron level. Our analysis is based on two popular and related methods in psychophysics and neurophysiology namely classification images and spike triggered analysis. These methods have been widely used to understand the underlying mechanisms of sensory systems in humans and monkeys. We leverage them to investigate the inherent biases of deep neural networks and to obtain a first-order approximation of their functionality. We emphasize on CNNs since they are currently the state of the art methods in computer vision and are a decent model of human visual processing. In addition, we study multi-layer perceptrons, logistic regression, and recurrent neural networks. Experiments over four classic datasets, MNIST, Fashion-MNIST, CIFAR-10, and ImageNet, show that the computed bias maps resemble the target classes and when used for classification lead to an over two-fold performance than the chance level. Further, we show that classification images can be used to attack a black-box classifier and to detect adversarial patch attacks. Finally, we utilize spike triggered averaging to derive the filters of CNNs and explore how the behavior of a network changes when neurons in different layers are modulated. Our effort illustrates a successful example of borrowing from neurosciences to study ANNs and highlights the importance of cross-fertilization and synergy across machine learning, deep learning, and computational neuroscience.

Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring

- Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, Jason Weston
- abstract@open-review(Poster): The use of deep pre-trained transformers has led to remarkable progress in a number of applications (Devlin et al., 2018). For tasks that make pairwise comparisons between sequences, matching a given input with a corresponding label, two approaches are common: Cross-encoders performing full self-attention over the pair and Bi-encoders encoding the pair separately. The former often performs better, but is too slow for practical use. In this work, we develop a new transformer architecture, the Poly-encoder, that learns global rather than token level self-attention features. We perform a detailed comparison of all three approaches, including what pre-training and fine-tuning strategies work best. We show our models achieve state-of-the-art results on four tasks; that Poly-encoders are faster than Cross-encoders and more accurate than Bi-encoders; and that the best results are obtained by pre-training on large datasets similar to the downstream tasks.

A Learning-based Iterative Method for Solving Vehicle Routing Problems

- Hao Lu, Xingwen Zhang, Shuang Yang
- abstract@open-review(Poster): This paper is concerned with solving combinatorial optimization problems, in particular, the capacitated vehicle routing problems (CVRP). Classical Operations Research (OR) algorithms such as LKH3 \citep{helsgaun2017extension} are inefficient and difficult to scale to larger-size problems. Machine learning based approaches have recently shown to be promising, partly because of their efficiency (once trained, they can perform solving within minutes or even seconds). However, there is still a considerable gap between the quality of a machine learned solution and what OR methods can offer (e.g., on CVRP-100, the best result of learned solutions is between 16.10-16.80, significantly worse than LKH3's 15.65). In this paper, we present ``Learn to Improve'' (L2I), the first learning based approach for CVRP that is efficient in solving speed and at the same time outperforms OR methods. Starting with a random initial solution, L2I learns to iteratively refine the solution with an improvement operator, selected by a reinforcement learning based controller. The improvement operator is selected from a pool of powerful operators that are customized for routing problems. By combining the strengths of the two worlds, our approach achieves the new state-of-the-art results on CVRP, e.g., an average cost of 15.57 on CVRP-100.

Transferable Perturbations of Deep Feature Distributions

- Nathan Inkawhich, Kevin Liang, Lawrence Carin, Yiran Chen
- abstract@open-review(Poster): Almost all current adversarial attacks of CNN classifiers rely on information derived from the output layer of the network. This work presents a new adversarial attack based on the modeling and exploitation of class-wise and layer-wise deep feature distributions. We achieve state-of-the-art targeted blackbox transfer-based attack results for undefended ImageNet models. Further, we place a priority on explainability and interpretability of the attacking process. Our methodology affords an analysis of how adversarial attacks change the intermediate feature distributions of CNNs, as well as a measure of layer-wise and class-wise feature distributional separability/entanglement. We also conceptualize a transition from task/data-specific to model-specific features within a CNN architecture that directly impacts the transferability of adversarial examples.

Skip Connections Matter: On the Transferability of Adversarial Examples Generated with ResNets

- Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, Xingjun Ma
- abstract@open-review(Spotlight): Skip connections are an essential component of current state-of-the-art deep neural networks (DNNs) such as ResNet, WideResNet, DenseNet, and ResNeXt. Despite their huge success in building deeper and more powerful DNNs, we identify a surprising \emph{security weakness} of skip connections in this paper. Use of skip connections \textit{allows easier generation of highly transferable adversarial examples}. Specifically, in ResNet-like (with skip connections) neural networks, gradients can backpropagate through either skip connections or residual modules. We find that using more gradients from the skip connections rather than the residual modules according to a decay factor, allows one to craft adversarial

examples with high transferability. Our method is termed \emph{Skip Gradient Method} (SGM). We conduct comprehensive transfer attacks against state-of-the-art DNNs including ResNets, DenseNets, Inceptions, Inception-ResNet, Squeeze-and-Excitation Network (SENet) and robustly trained DNNs. We show that employing SGM on the gradient flow can greatly improve the transferability of crafted attacks in almost all cases. Furthermore, SGM can be easily combined with existing black-box attack techniques, and obtain high improvements over state-of-the-art transferability methods. Our findings not only motivate new research into the architectural vulnerability of DNNs, but also open up further challenges for the design of secure DNN architectures.

A Signal Propagation Perspective for Pruning Neural Networks at Initialization

- Namhoon Lee, Thalaiyasingam Ajanthan, Stephen Gould, Philip H. S. Torr
- abstract@open-review(Spotlight): Network pruning is a promising avenue for compressing deep neural networks. A typical approach to pruning starts by training a model and then removing redundant parameters while minimizing the impact on what is learned. Alternatively, a recent approach shows that pruning can be done at initialization prior to training, based on a saliency criterion called connection sensitivity. However, it remains unclear exactly why pruning an untrained, randomly initialized neural network is effective. In this work, by noting connection sensitivity as a form of gradient, we formally characterize initialization conditions to ensure reliable connection sensitivity measurements, which in turn yields effective pruning results. Moreover, we analyze the signal propagation properties of the resulting pruned networks and introduce a simple, data-free method to improve their trainability. Our modifications to the existing pruning at initialization method lead to improved results on all tested network models for image classification tasks. Furthermore, we empirically study the effect of supervision for pruning and demonstrate that our signal propagation perspective, combined with unsupervised pruning, can be useful in various scenarios where pruning is applied to non-standard arbitrarily-designed architectures.

Continual Learning with Adaptive Weights (CLAW)

- Tameem Adel, Han Zhao, Richard E. Turner
- abstract@open-review(Poster): Approaches to continual learning aim to successfully learn a set of related tasks that arrive in an online manner. Recently, several frameworks have been developed which enable deep learning to be deployed in this learning scenario. A key modelling decision is to what extent the architecture should be shared across tasks. On the one hand, separately modelling each task avoids catastrophic forgetting but it does not support transfer learning and leads to large models. On the other hand, rigidly specifying a shared component and a task-specific part enables task transfer and limits the model size, but it is vulnerable to catastrophic forgetting and restricts the form of task-transfer that can occur. Ideally, the network should adaptively identify which parts of the network to share in a data driven way. Here we introduce such an approach called Continual Learning with Adaptive Weights (CLAW), which is based on probabilistic modelling and variational inference. Experiments show that CLAW achieves state-of-the-art performance on six benchmarks in terms of overall continual learning performance, as measured by classification accuracy, and in terms of addressing catastrophic forgetting.

Intensity-Free Learning of Temporal Point Processes

- Oleksandr Shchur, Marin Biloš, Stephan Günnemann
- abstract@open-review(Spotlight): Temporal point processes are the dominant paradigm for modeling sequences of events happening at irregular intervals. The standard way of learning in such models is by estimating the conditional intensity function. However, parameterizing the intensity function usually incurs several trade-offs. We show how to overcome the limitations of intensity-based approaches by directly modeling the conditional distribution of inter-event times. We draw on the literature on normalizing flows to design models that are flexible and efficient. We additionally propose a simple mixture model that matches the flexibility of flow-based models, but also permits sampling and computing moments in closed form. The proposed models achieve state-of-the-art performance in standard prediction tasks and are suitable for novel applications, such as learning sequence embeddings and imputing missing data.

Scalable and Order-robust Continual Learning with Additive Parameter Decomposition

- Jaehong Yoon, Saehoon Kim, Eunho Yang, Sung Ju Hwang
- abstract@open-review(Poster): While recent continual learning methods largely alleviate the catastrophic problem on toy-sized datasets, there are issues that remain to be tackled in order to apply them to real-world problem domains. First, a continual learning model should effectively handle catastrophic forgetting and be efficient to train even with a large number of tasks. Secondly, it needs to tackle the problem of order-sensitivity, where the performance of the tasks largely varies based on the order of the task arrival sequence, as it may cause serious problems where fairness plays a critical role (e.g. medical diagnosis). To tackle these practical challenges, we propose a novel continual learning method that is scalable as well as order-robust, which instead of learning a completely shared set of weights, represents the parameters for each task as a sum of task-shared and sparse task-adaptive parameters. With our Additive Parameter Decomposition (APD), the task-adaptive parameters for earlier tasks remain mostly unaffected, where we update them only to reflect the changes made to the task-shared parameters. This decomposition of parameters effectively prevents catastrophic forgetting and order-sensitivity, while being computation- and memory-efficient. Further, we can achieve even better scalability with APD using hierarchical knowledge consolidation, which clusters the task-adaptive parameters to obtain hierarchically shared parameters. We validate our network with APD, APD-Net, on multiple benchmark datasets against state-of-the-art continual learning methods, which it largely outperforms in accuracy, scalability, and order-robustness.

Weakly Supervised Clustering by Exploiting Unique Class Count

- Mustafa Umit Oner, Hwee Kuan Lee, Wing-Kin Sung
- abstract@open-review(Poster): A weakly supervised learning based clustering framework is proposed in this paper. As the core of this framework, we introduce a novel multiple instance learning task based on a bag level label called unique class count (ucc), which is the number of unique classes among all instances inside the bag. In this task, no annotations on individual instances inside the bag are needed during training of the models. We mathematically prove that with a perfect ucc classifier, perfect clustering of individual instances inside the bags is possible even when no annotations on individual instances are given during training. We have constructed a neural network based ucc classifier and experimentally shown that the clustering performance of our framework with our weakly supervised ucc classifier is comparable to that of fully supervised learning models where labels for all instances are known. Furthermore, we have tested the applicability of our framework to a real world task of semantic segmentation of breast cancer metastases in histological lymph node sections and shown that the performance of our weakly supervised framework is comparable to the performance of a fully supervised Unet model.

Learning to Control PDEs with Differentiable Physics

- Philipp Holl, Nils Thuerey, Vladlen Koltun
- abstract@open-review(Spotlight): Predicting outcomes and planning interactions with the physical world are long-standing goals for machine learning. A variety of such tasks involves continuous physical systems, which can be described by partial differential equations (PDEs) with many degrees of freedom. Existing methods that aim to control the dynamics of such systems are typically limited to relatively short time frames or a small number of interaction parameters. We present a novel hierarchical predictor-corrector scheme which enables neural networks to learn to understand and control complex nonlinear physical systems over long time frames. We propose to split the problem into two distinct tasks: planning and control. To this end, we introduce a predictor network that plans optimal trajectories and a control network that infers the corresponding control parameters. Both stages are

trained end-to-end using a differentiable PDE solver. We demonstrate that our method successfully develops an understanding of complex physical systems and learns to control them for tasks involving PDEs such as the incompressible Navier-Stokes equations.

Linear Symmetric Quantization of Neural Networks for Low-precision Integer Hardware

- Xiandong Zhao, Ying Wang, Xuyi Cai, Cheng Liu, Lei Zhang
- abstract@open-review(Poster): With the proliferation of specialized neural network processors that operate on low-precision integers, the performance of Deep Neural Network inference becomes increasingly dependent on the result of quantization. Despite plenty of prior work on the quantization of weights or activations for neural networks, there is still a wide gap between the software quantizers and the low-precision accelerator implementation, which degrades either the efficiency of networks or that of the hardware for the lack of software and hardware coordination at design-phase. In this paper, we propose a learned linear symmetric quantizer for integer neural network processors, which not only quantizes neural parameters and activations to low-bit integer but also accelerates hardware inference by using batch normalization fusion and low-precision accumulators (e.g., 16-bit) and multipliers (e.g., 4-bit). We use a unified way to quantize weights and activations, and the results outperform many previous approaches for various networks such as AlexNet, ResNet, and lightweight models like MobileNet while keeping friendly to the accelerator architecture. Additional, we also apply the method to object detection models and witness high performance and accuracy in YOLO-v2. Finally, we deploy the quantized models on our specialized integer-arithmetic-only DNN accelerator to show the effectiveness of the proposed quantizer. We show that even with linear symmetric quantization, the results can be better than asymmetric or non-linear methods in 4-bit networks. In evaluation, the proposed quantizer induces less than 0.4\% accuracy drop in ResNet18, ResNet34, and AlexNet when quantizing the whole network as required by the integer processors.

Estimating Gradients for Discrete Random Variables by Sampling without Replacement

- Wouter Kool, Herke van Hoof, Max Welling
- abstract@open-review(Spotlight): We derive an unbiased estimator for expectations over discrete random variables based on sampling without replacement, which reduces variance as it avoids duplicate samples. We show that our estimator can be derived as the Rao-Blackwellization of three different estimators. Combining our estimator with REINFORCE, we obtain a policy gradient estimator and we reduce its variance using a built-in control variate which is obtained without additional model evaluations. The resulting estimator is closely related to other gradient estimators. Experiments with a toy problem, a categorical Variational Auto-Encoder and a structured prediction problem show that our estimator is the only estimator that is consistently among the best estimators in both high and low entropy settings.

To Relieve Your Headache of Training an MRF, Take AdVIL

- Chongxuan Li, Chao Du, Kun Xu, Max Welling, Jun Zhu, Bo Zhang
- abstract@open-review(Poster): We propose a black-box algorithm called {\it Adversarial Variational Inference and Learning} (AdVIL) to perform inference and learning on a general Markov random field (MRF). AdVIL employs two variational distributions to approximately infer the latent variables and estimate the partition function of an MRF, respectively. The two variational distributions provide an estimate of the negative log-likelihood of the MRF as a minimax optimization problem, which is solved by stochastic gradient descent. AdVIL is proven convergent under certain conditions. On one hand, compared with contrastive divergence, AdVIL requires a minimal assumption about the model structure and can deal with a broader family of MRFs. On the other hand, compared with existing black-box methods, AdVIL provides a tighter estimate of the log partition function and achieves much better empirical results.

Automated Relational Meta-learning

- Huaxiu Yao, Xian Wu, Zhiqiang Tao, Yaliang Li, Bolin Ding, Ruirui Li, Zhenhui Li
- abstract@open-review(Poster): In order to efficiently learn with small amount of data on new tasks, meta-learning transfers knowledge learned from previous tasks to the new ones. However, a critical challenge in meta-learning is the task heterogeneity which cannot be well handled by traditional globally shared meta-learning methods. In addition, current task-specific meta-learning methods may either suffer from hand-crafted structure design or lack the capability to capture complex relations between tasks. In this paper, motivated by the way of knowledge organization in knowledge bases, we propose an automated relational meta-learning (ARML) framework that automatically extracts the cross-task relations and constructs the meta-knowledge graph. When a new task arrives, it can quickly find the most relevant structure and tailor the learned structure knowledge to the meta-learner. As a result, the proposed framework not only addresses the challenge of task heterogeneity by a learned meta-knowledge graph, but also increases the model interpretability. We conduct extensive experiments on 2D toy regression and few-shot image classification and the results demonstrate the superiority of ARML over state-of-the-art baselines.

Defending Against Physically Realizable Attacks on Image Classification

- Tong Wu, Liang Tong, Yevgeniy Vorobeychik
- abstract@open-review(Spotlight): We study the problem of defending deep neural network approaches for image classification from physically realizable attacks. First, we demonstrate that the two most scalable and effective methods for learning robust models, adversarial training with PGD attacks and randomized smoothing, exhibit very limited effectiveness against three of the highest profile physical attacks. Next, we propose a new abstract adversarial model, rectangular occlusion attacks, in which an adversary places a small adversarially crafted rectangle in an image, and develop two approaches for efficiently computing the resulting adversarial examples. Finally, we demonstrate that adversarial training using our new attack yields image classification models that exhibit high robustness against the physically realizable attacks we study, offering the first effective generic defense against such attacks.

N-BEATS: Neural basis expansion analysis for interpretable time series forecasting

- Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, Yoshua Bengio
- abstract@open-review(Poster): We focus on solving the univariate times series point forecasting problem using deep learning. We propose a deep neural architecture based on backward and forward residual links and a very deep stack of fully-connected layers. The architecture has a number of desirable properties, being interpretable, applicable without modification to a wide array of target domains, and fast to train. We test the proposed architecture on several well-known datasets, including M3, M4 and TOURISM competition datasets containing time series from diverse domains. We demonstrate state-of-the-art performance for two configurations of N-BEATS for all the datasets, improving forecast accuracy by 11% over a statistical benchmark and by 3% over last year's winner of the M4 competition, a domain-adjusted hand-crafted hybrid between neural network and statistical time series models. The first configuration of our model does not employ any time-series-specific components and its performance on heterogeneous datasets strongly suggests that, contrarily to received wisdom, deep learning primitives such as residual blocks are by themselves sufficient to solve a wide range of forecasting problems. Finally, we demonstrate how the proposed architecture can be augmented to provide outputs that are interpretable without considerable loss in accuracy.

Deep Learning of Determinantal Point Processes via Proper Spectral Sub-gradient

- Tianshu Yu, Yikang Li, Baoxin Li

- abstract@open-review(Poster): Determinantal point processes (DPPs) is an effective tool to deliver diversity on multiple machine learning and computer vision tasks. Under deep learning framework, DPP is typically optimized via approximation, which is not straightforward and has some conflict with diversity requirement. We note, however, there has been no deep learning paradigms to optimize DPP directly since it involves matrix inversion which may result in highly computational instability. This fact greatly hinders the wide use of DPP on some specific objectives where DPP serves as a term to measure the feature diversity. In this paper, we devise a simple but effective algorithm to address this issue to optimize DPP term directly expressed with L-ensemble in spectral domain over gram matrix, which is more flexible than learning on parametric kernels. By further taking into account some geometric constraints, our algorithm seeks to generate valid sub-gradients of DPP term in case when the DPP gram matrix is not invertible (no gradients exist in this case). In this sense, our algorithm can be easily incorporated with multiple deep learning tasks. Experiments show the effectiveness of our algorithm, indicating promising performance for practical learning problems.

Distance-Based Learning from Errors for Confidence Calibration

- Chen Xing, Sercan Arik, Zizhao Zhang, Tomas Pfister
- abstract@open-review(Poster): Deep neural networks (DNNs) are poorly calibrated when trained in conventional ways. To improve confidence calibration of DNNs, we propose a novel training method, distance-based learning from errors (DBLE). DBLE bases its confidence estimation on distances in the representation space. In DBLE, we first adapt prototypical learning to train classification models. It yields a representation space where the distance between a test sample and its ground truth class center can calibrate the model's classification performance. At inference, however, these distances are not available due to the lack of ground truth labels. To circumvent this by inferring the distance for every test sample, we propose to train a confidence model jointly with the classification model. We integrate this into training by merely learning from mis-classified training samples, which we show to be highly beneficial for effective learning. On multiple datasets and DNN architectures, we demonstrate that DBLE outperforms alternative single-model confidence calibration approaches. DBLE also achieves comparable performance with computationally-expensive ensemble approaches with lower computational cost and lower number of parameters.

Curvature Graph Network

- Ze Ye, Kin Sum Liu, Tengfei Ma, Jie Gao, Chao Chen
- abstract@open-review(Poster): Graph-structured data is prevalent in many domains. Despite the widely celebrated success of deep neural networks, their power in graph-structured data is yet to be fully explored. We propose a novel network architecture that incorporates advanced graph structural features. In particular, we leverage discrete graph curvature, which measures how the neighborhoods of a pair of nodes are structurally related. The curvature of an edge (x, y) defines the distance taken to travel from neighbors of x to neighbors of y, compared with the length of edge (x, y). It is a much more descriptive feature compared to previously used features that only focus on node specific attributes or limited topological information such as degree. Our curvature graph convolution network outperforms state-of-the-art on various synthetic and real-world graphs, especially the larger and denser ones.

Learning Expensive Coordination: An Event-Based Deep RL Approach

- Zhenyu Shi, *Runsheng Yu*, Xinrun Wang*, Rundong Wang, Youzhi Zhang, Hanjiang Lai, Bo An
- abstract@open-review(Poster): Existing works in deep Multi-Agent Reinforcement Learning (MARL) mainly focus on coordinating cooperative agents to complete certain tasks jointly. However, in many cases of the real world, agents are self-interested such as employees in a company and clubs in a league. Therefore, the leader, i.e., the manager of the company or the league, needs to provide bonuses to followers for efficient coordination, which we call expensive coordination. The main difficulties of expensive coordination are that i) the leader has to consider the long-term effect and predict the followers' behaviors when assigning bonuses and ii) the complex interactions between followers make the training process hard to converge, especially when the leader's policy changes with time. In this work, we address this problem through an event-based deep RL approach. Our main contributions are threefold. (1) We model the leader's decision-making process as a semi-Markov Decision Process and propose a novel multi-agent event-based policy gradient to learn the leader's long-term policy. (2) We exploit the leader-follower consistency scheme to design a follower-aware module and a follower-specific attention module to predict the followers' behaviors and make accurate response to their behaviors. (3) We propose an action abstraction-based policy gradient algorithm to reduce the followers' decision space and thus accelerate the training process of followers. Experiments in resource collections, navigation, and the predator-prey game reveal that our approach outperforms the state-of-the-art methods dramatically.

LAMOL: LAnguage MOdeling for Lifelong Language Learning

- Fan-Keng Sun, *Cheng-Hao Ho*, Hung-Yi Lee
- abstract@open-review(Poster): Most research on lifelong learning applies to images or games, but not language. We present LAMOL, a simple yet effective method for lifelong language learning (LLL) based on language modeling. LAMOL replays pseudo-samples of previous tasks while requiring no extra memory or model capacity. Specifically, LAMOL is a language model that simultaneously learns to solve the tasks and generate training samples. When the model is trained for a new task, it generates pseudo-samples of previous tasks for training alongside data for the new task. The results show that LAMOL prevents catastrophic forgetting without any sign of intransigence and can perform five very different language tasks sequentially with only one model. Overall, LAMOL outperforms previous methods by a considerable margin and is only 2-3% worse than multitasking, which is usually considered the LLL upper bound. The source code is available at https://github.com/jojotenya/LAMOL.

GenDICE: Generalized Offline Estimation of Stationary Values

- Ruiyi Zhang, *Bo Dai*, Lihong Li, Dale Schuurmans
- abstract@open-review(Talk): An important problem that arises in reinforcement learning and Monte Carlo methods is estimating quantities defined by the stationary distribution of a Markov chain. In many real-world applications, access to the underlying transition operator is limited to a fixed set of data that has already been collected, without additional interaction with the environment being available. We show that consistent estimation remains possible in this scenario, and that effective estimation can still be achieved in important applications. Our approach is based on estimating a ratio that corrects for the discrepancy between the stationary and empirical distributions, derived from fundamental properties of the stationary distribution, and exploiting constraint reformulations based on variational divergence minimization. The resulting algorithm, GenDICE, is straightforward and effective. We prove the consistency of the method under general conditions, provide a detailed error analysis, and demonstrate strong empirical performance on benchmark tasks, including off-line PageRank and off-policy policy evaluation.

ProxSGD: Training Structured Neural Networks under Regularization and Constraints

- Yang Yang, Yaxiong Yuan, Avraam Chatzimichailidis, Ruud JG van Sloun, Lei Lei, Symeon Chatzinotas
- abstract@open-review(Poster): In this paper, we consider the problem of training neural networks (NN). To promote a NN with specific structures, we explicitly take into consideration the nonsmooth regularization (such as L1-norm) and constraints (such as interval constraint). This is formulated as a constrained nonsmooth nonconvex optimization problem, and we propose a convergent proximal-type stochastic gradient descent (Prox-SGD) algorithm. We show that under properly selected learning rates, momentum eventually resembles the unknown real gradient and thus is crucial in analyzing the convergence. We establish that with probability 1, every limit point of the sequence generated by the proposed Prox-SGD is a stationary point. Then the Prox-SGD is tailored to train a sparse neural network and a binary neural network, and the theoretical analysis is also supported by extensive numerical tests.

Diverse Trajectory Forecasting with Determinantal Point Processes

- Ye Yuan, Kris M. Kitani
- abstract@open-review(Poster): The ability to forecast a set of likely yet diverse possible future behaviors of an agent (e.g., future trajectories of a pedestrian) is essential for safety-critical perception systems (e.g., autonomous vehicles). In particular, a set of possible future behaviors generated by the system must be diverse to account for all possible outcomes in order to take necessary safety precautions. It is not sufficient to maintain a set of the most likely future outcomes because the set may only contain perturbations of a dominating single outcome (major mode). While generative models such as variational autoencoders (VAEs) have been shown to be a powerful tool for learning a distribution over future trajectories, randomly drawn samples from the learned implicit likelihood model may not be diverse -- the likelihood model is derived from the training data distribution and the samples will concentrate around the major mode of the data. In this work, we propose to learn a diversity sampling function (DSF) that generates a diverse yet likely set of future trajectories. The DSF maps forecasting context features to a set of latent codes which can be decoded by a generative model (e.g., VAE) into a set of diverse trajectory samples. Concretely, the process of identifying the diverse set of samples is posed as DSF parameter estimation. To learn the parameters of the DSF, the diversity of the trajectory samples is evaluated by a diversity loss based on a determinantal point process (DPP). Gradient descent is performed over the DSF parameters, which in turn moves the latent codes of the sample set to find an optimal set of diverse yet likely trajectories. Our method is a novel application of DPPs to optimize a set of items (forecasted trajectories) in continuous space. We demonstrate the diversity of the trajectories produced by our approach on both low-dimensional 2D trajectory data and high-dimensional human motion data.

Evaluating The Search Phase of Neural Architecture Search

- Kaicheng Yu, Christian Sciuto, Martin Jaggi, Claudiu Musat, Mathieu Salzmann
- abstract@open-review(Poster): Neural Architecture Search (NAS) aims to facilitate the design of deep networks for new tasks. Existing techniques rely on two stages: searching over the architecture space and validating the best architecture. NAS algorithms are currently compared solely based on their results on the downstream task. While intuitive, this fails to explicitly evaluate the effectiveness of their search strategies. In this paper, we propose to evaluate the NAS search phase. To this end, we compare the quality of the solutions obtained by NAS search policies with that of random architecture selection. We find that: (i) On average, the state-of-the-art NAS algorithms perform similarly to the random policy; (ii) the widely-used weight sharing strategy degrades the ranking of the NAS candidates to the point of not reflecting their true performance, thus reducing the effectiveness of the search process. We believe that our evaluation framework will be key to designing NAS strategies that consistently discover architectures superior to random ones.

On Robustness of Neural Ordinary Differential Equations

- Hanshu YAN, Jiawei DU, Vincent TAN, Jiashi FENG
- abstract@open-review(Spotlight): Neural ordinary differential equations (ODEs) have been attracting increasing attention in various research domains recently. There have been some works studying optimization issues and approximation capabilities of neural ODEs, but their robustness is still yet unclear. In this work, we fill this important gap by exploring robustness properties of neural ODEs both empirically and theoretically. We first present an empirical study on the robustness of the neural ODE-based networks (ODENets) by exposing them to inputs with various types of perturbations and subsequently investigating the changes of the corresponding outputs. In contrast to conventional convolutional neural networks (CNNs), we find that the ODENets are more robust against both random Gaussian perturbations and adversarial attack examples. We then provide an insightful understanding of this phenomenon by exploiting a certain desirable property of the flow of a continuous-time ODE, namely that integral curves are non-intersecting. Our work suggests that, due to their intrinsic robustness, it is promising to use neural ODEs as a basic block for building robust deep network models. To further enhance the robustness of vanilla neural ODEs, we propose the time-invariant steady neural ODE (TisODE), which regularizes the flow on perturbed data via the time-invariant property and the imposition of a steady-state constraint. We show that the TisODE method outperforms vanilla neural ODEs and also can work in conjunction with other state-of-the-art architectural methods to build more robust deep networks.

BackPACK: Packing more into Backprop

- Felix Dangel, Frederik Kunstner, Philipp Hennig
- abstract@open-review(Talk): Automatic differentiation frameworks are optimized for exactly one thing: computing the average mini-batch gradient. Yet, other quantities such as the variance of the mini-batch gradients or many approximations to the Hessian can, in theory, be computed efficiently, and at the same time as the gradient. While these quantities are of great interest to researchers and practitioners, current deep learning software does not support their automatic calculation. Manually implementing them is burdensome, inefficient if done naively, and the resulting code is rarely shared. This hampers progress in deep learning, and unnecessarily narrows research to focus on gradient descent and its variants; it also complicates replication studies and comparisons between newly developed methods that require those quantities, to the point of impossibility. To address this problem, we introduce BackPACK, an efficient framework built on top of PyTorch, that extends the backpropagation algorithm to extract additional information from first-and second-order derivatives. Its capabilities are illustrated by benchmark reports for computing additional quantities on deep neural networks, and an example application by testing several recent curvature approximations for optimization.

DeepHoyer: Learning Sparser Neural Network with Differentiable Scale-Invariant Sparsity Measures

- Huanrui Yang, Wei Wen, Hai Li
- abstract@open-review(Poster): In seeking for sparse and efficient neural network models, many previous works investigated on enforcing L1 or L0 regularizers to encourage weight sparsity during training. The L0 regularizer measures the parameter sparsity directly and is invariant to the scaling of parameter values. But it cannot provide useful gradients and therefore requires complex optimization techniques. The L1 regularizer is almost everywhere differentiable and can be easily optimized with gradient descent. Yet it is not scale-invariant and causes the same shrinking rate to all parameters, which is inefficient in increasing sparsity. Inspired by the Hoyer measure (the ratio between L1 and L2 norms) used in traditional compressed sensing problems, we present DeepHoyer, a set of sparsity-inducing regularizers that are both differentiable almost everywhere and scale-invariant. Our experiments show that enforcing DeepHoyer regularizers can produce even sparser neural network models than previous works, under the same accuracy level. We also show that DeepHoyer can be applied to both element-wise and structural pruning.

Depth-Adaptive Transformer

- Maha Elbayad, Jiatao Gu, Edouard Grave, Michael Auli
- abstract@open-review(Poster): State of the art sequence-to-sequence models for large scale tasks perform a fixed number of computations for each input sequence regardless of whether it is easy or hard to process. In this paper, we train Transformer models which can make output predictions at different stages of the network and we investigate different ways to predict how much computation is required for a particular sequence. Unlike dynamic computation in Universal Transformers, which applies the same set of layers iteratively, we apply different layers at every step to adjust both the amount of computation as well as the model capacity. On IWSLT German-English translation our approach matches the accuracy of a well tuned baseline Transformer while using less than a quarter of the decoder layers.

InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization

- Fan-Yun Sun, Jordan Hoffman, Vikas Verma, Jian Tang

- abstract@open-review(Spotlight): This paper studies learning the representations of whole graphs in both unsupervised and semi-supervised scenarios. Graph-level representations are critical in a variety of real-world applications such as predicting the properties of molecules and community analysis in social networks. Traditional graph kernel based methods are simple, yet effective for obtaining fixed-length representations for graphs but they suffer from poor generalization due to hand-crafted designs. There are also some recent methods based on language models (e.g. graph2vec) but they tend to only consider certain substructures (e.g. subtrees) as graph representatives. Inspired by recent progress of unsupervised representation learning, in this paper we proposed a novel method called InfoGraph for learning graph-level representations. We maximize the mutual information between the graph-level representation and the representations of substructures of different scales (e.g., nodes, edges, triangles). By doing so, the graph-level representations encode aspects of the data that are shared across different scales of substructures. Furthermore, we further propose InfoGraph*, an extension of InfoGraph for semisupervised scenarios*. *InfoGraph* maximizes the mutual information between unsupervised graph representations learned by InfoGraph and the representations learned by existing supervised methods. As a result, the supervised encoder learns from unlabeled data while preserving the latent semantic space favored by the current supervised task. Experimental results on the tasks of graph classification and molecular property prediction show that InfoGraph is superior to state-of-the-art baselines and InfoGraph* can achieve performance competitive with state-of-the-art semi-supervised models.

Federated Adversarial Domain Adaptation

- Xingchao Peng, Zijun Huang, Yizhe Zhu, Kate Saenko
- abstract@open-review(Poster): Federated learning improves data privacy and efficiency in machine learning performed over networks of distributed devices, such as mobile phones, IoT and wearable devices, etc. Yet models trained with federated learning can still fail to generalize to new devices due to the problem of domain shift. Domain shift occurs when the labeled data collected by source nodes statistically differs from the target node's unlabeled data. In this work, we present a principled approach to the problem of federated domain adaptation, which aims to align the representations learned among the different nodes with the data distribution of the target node. Our approach extends adversarial adaptation techniques to the constraints of the federated setting. In addition, we devise a dynamic attention mechanism and leverage feature disentanglement to enhance knowledge transfer. Empirically, we perform extensive experiments on several image and text classification tasks and show promising results under unsupervised federated domain adaptation setting.

CATER: A diagnostic dataset for Compositional Actions & TEmporal Reasoning

- Rohit Girdhar, Deva Ramanan
- abstract@open-review(Talk): Computer vision has undergone a dramatic revolution in performance, driven in large part through deep features trained on large-scale supervised datasets. However, much of these improvements have focused on static image analysis; video understanding has seen rather modest improvements. Even though new datasets and spatiotemporal models have been proposed, simple frame-by-frame classification methods often still remain competitive. We posit that current video datasets are plagued with implicit biases over scene and object structure that can dwarf variations in temporal structure. In this work, we build a video dataset with fully observable and controllable object and scene bias, and which truly requires spatiotemporal understanding in order to be solved. Our dataset, named CATER, is rendered synthetically using a library of standard 3D objects, and tests the ability to recognize compositions of object movements that require long-term reasoning. In addition to being a challenging dataset, CATER also provides a plethora of diagnostic tools to analyze modern spatiotemporal video architectures by being completely observable and controllable. Using CATER, we provide insights into some of the most recent state of the art deep video architectures.

Maxmin Q-learning: Controlling the Estimation Bias of Q-learning

- Qingfeng Lan, Yangchen Pan, Alona Fyshe, Martha White
- abstract@open-review(Poster): Q-learning suffers from overestimation bias, because it approximates the maximum action value using the maximum estimated action value. Algorithms have been proposed to reduce overestimation bias, but we lack an understanding of how bias interacts with performance, and the extent to which existing algorithms mitigate bias. In this paper, we 1) highlight that the effect of overestimation bias on learning efficiency is environment-dependent; 2) propose a generalization of Q-learning, called \emph{Maxmin Q-learning}, which provides a parameter to flexibly control bias; 3) show theoretically that there exists a parameter choice for Maxmin Q-learning that leads to unbiased estimation with a lower approximation variance than Q-learning; and 4) prove the convergence of our algorithm in the tabular case, as well as convergence of several previous Q-learning variants, using a novel Generalized Q-learning framework. We empirically verify that our algorithm better controls estimation bias in toy environments, and that it achieves superior performance on several benchmark problems.

Automatically Discovering and Learning New Visual Categories with Ranking Statistics

- Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, Andrew Zisserman
- abstract@open-review(Poster): We tackle the problem of discovering novel classes in an image collection given labelled examples of other classes. This setting is similar to semi-supervised learning, but significantly harder because there are no labelled examples for the new classes. The challenge, then, is to leverage the information contained in the labelled images in order to learn a general-purpose clustering model and use the latter to identify the new classes in the unlabelled data. In this work we address this problem by combining three ideas: (1) we suggest that the common approach of bootstrapping an image representation using the labeled data only introduces an unwanted bias, and that this can be avoided by using self-supervised learning to train the representation from scratch on the union of labelled and unlabelled data; (2) we use rank statistics to transfer the model's knowledge of the labelled classes to the problem of clustering the unlabelled images; and, (3) we train the data representation by optimizing a joint objective function on the labelled and unlabelled subsets of the data, improving both the supervised classification of the labelled data, and the clustering of the unlabelled data. We evaluate our approach on standard classification benchmarks and outperform current methods for novel category discovery by a significant margin.

Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-identification

- Yixiao Ge, Dapeng Chen, Hongsheng Li
- abstract@open-review(Poster): Person re-identification (re-ID) aims at identifying the same persons' images across different cameras. However, domain diversities between different datasets pose an evident challenge for adapting the re-ID model trained on one dataset to another one. State-of-the-art unsupervised domain adaptation methods for person re-ID transferred the learned knowledge from the source domain by optimizing with pseudo labels created by clustering algorithms on the target domain. Although they achieved state-of-the-art performances, the inevitable label noise caused by the clustering procedure was ignored. Such noisy pseudo labels substantially hinders the model's capability on further improving feature representations on the target domain. In order to mitigate the effects of noisy pseudo labels, we propose to softly refine the pseudo labels in the target domain by proposing an unsupervised framework, Mutual Mean-Teaching (MMT), to learn better features from the target domain via off-line refined hard pseudo labels and on-line refined soft pseudo labels in an alternative training manner. In addition, the common practice is to adopt both the classification loss and the triplet loss jointly for achieving optimal performances in person re-ID models. However, conventional triplet loss cannot work with softly refined labels. To solve this problem, a novel soft softmax-triplet loss is proposed to support learning with soft pseudo triplet labels for achieving the optimal domain adaptation performance. The proposed MMT framework achieves considerable improvements of 14.4%, 18.2%, 13.1% and 16.4% mAP on Market-to-Duke, Duke-to-Market, Market-to-MSMT and Duke-to-MSMT unsupervised domain adaptation tasks.

Multi-Scale Representation Learning for Spatial Feature Distributions using Grid Cells

- Gengchen Mai, Krzysztof Janowicz, Bo Yan, Rui Zhu, Ling Cai, Ni Lao

- abstract@open-review(Spotlight): Unsupervised text encoding models have recently fueled substantial progress in NLP. The key idea is to use neural networks to convert words in texts to vector space representations (embeddings) based on word positions in a sentence and their contexts, which are suitable for end-to-end training of downstream tasks. We see a strikingly similar situation in spatial analysis, which focuses on incorporating both absolute positions and spatial contexts of geographic objects such as POIs into models. A general-purpose representation model for space is valuable for a multitude of tasks. However, no such general model exists to date beyond simply applying discretization or feed-forward nets to coordinates, and little effort has been put into jointly modeling distributions with vastly different characteristics, which commonly emerges from GIS data. Meanwhile, Nobel Prize-winning Neuroscience research shows that grid cells in mammals provide a multi-scale periodic representation that functions as a metric for location encoding and is critical for recognizing places and for path-integration. Therefore, we propose a representation learning model called Space2Vec to encode the absolute positions and spatial relationships of places. We conduct experiments on two real-world geographic data for two different tasks: 1) predicting types of POIs given their positions and context, 2) image classification leveraging their geo-locations. Results show that because of its multi-scale representations, Space2Vec outperforms well-established ML approaches such as RBF kernels, multi-layer feed-forward nets, and tile embedding approaches for location modeling and image classification tasks. Detailed analysis shows that all baselines can at most well handle distribution at one scale but show poor performances in other scales. In contrast, Space2Vec 's multi-scale representation can handle distributions at different scales.

Neural Oblivious Decision Ensembles for Deep Learning on Tabular Data

- Sergei Popov, Stanislav Morozov, Artem Babenko
- abstract@open-review(Poster): Nowadays, deep neural networks (DNNs) have become the main instrument for machine learning tasks within a wide range of domains, including vision, NLP, and speech. Meanwhile, in an important case of heterogenous tabular data, the advantage of DNNs over shallow counterparts remains questionable. In particular, there is no sufficient evidence that deep learning machinery allows constructing methods that outperform gradient boosting decision trees (GBDT), which are often the top choice for tabular problems. In this paper, we introduce Neural Oblivious Decision Ensembles (NODE), a new deep learning architecture, designed to work with any tabular data. In a nutshell, the proposed NODE architecture generalizes ensembles of oblivious decision trees, but benefits from both end-to-end gradient-based optimization and the power of multi-layer hierarchical representation learning. With an extensive experimental comparison to the leading GBDT packages on a large number of tabular datasets, we demonstrate the advantage of the proposed NODE architecture, which outperforms the competitors on most of the tasks. We open-source the PyTorch implementation of NODE and believe that it will become a universal framework for machine learning on tabular data.

SQIL: Imitation Learning via Reinforcement Learning with Sparse Rewards

- Siddharth Reddy, Anca D. Dragan, Sergey Levine
- abstract@open-review(Poster): Learning to imitate expert behavior from demonstrations can be challenging, especially in environments with high-dimensional, continuous observations and unknown dynamics. Supervised learning methods based on behavioral cloning (BC) suffer from distribution shift: because the agent greedily imitates demonstrated actions, it can drift away from demonstrated states due to error accumulation. Recent methods based on reinforcement learning (RL), such as inverse RL and generative adversarial imitation learning (GAIL), overcome this issue by training an RL agent to match the demonstrations over a long horizon. Since the true reward function for the task is unknown, these methods learn a reward function from the demonstrations, often using complex and brittle approximation techniques that involve adversarial training. We propose a simple alternative that still uses RL, but does not require learning a reward function. The key idea is to provide the agent with an incentive to match the demonstrations over a long horizon, by encouraging it to return to demonstrated states upon encountering new, out-of-distribution states. We accomplish this by giving the agent a constant reward of r=+1 for matching the demonstrated action in a demonstrated state, and a constant reward of r=0 for all other behavior. Our method, which we call soft Q imitation learning (SQIL), can be implemented with a handful of minor modifications to any standard Q-learning or off-policy actor-critic algorithm. Theoretically, we show that SQIL can be interpreted as a regularized variant of BC that uses a sparsity prior to encourage long-horizon imitation. Empirically, we show that SQIL outperforms BC and achieves competitive results compared to GAIL, on a variety of image-based and low-dimensional tasks in Box2D, Atari, and MuJoCo. This paper is a proof of concept that illustrates how a simple imitation method based on RL with constant rewards can be as effective as more complex methods that use learned rewards.

Graph Neural Networks Exponentially Lose Expressive Power for Node Classification

- Kenta Oono, Taiji Suzuki
- abstract@open-review(Spotlight): Graph Neural Networks (graph NNs) are a promising deep learning approach for analyzing graph-structured data. However, it is known that they do not improve (or sometimes worsen) their predictive performance as we pile up many layers and add non-lineality. To tackle this problem, we investigate the expressive power of graph NNs via their asymptotic behaviors as the layer size tends to infinity. Our strategy is to generalize the forward propagation of a Graph Convolutional Network (GCN), which is a popular graph NN variant, as a specific dynamical system. In the case of a GCN, we show that when its weights satisfy the conditions determined by the spectra of the (augmented) normalized Laplacian, its output exponentially approaches the set of signals that carry information of the connected components and node degrees only for distinguishing nodes. Our theory enables us to relate the expressive power of GCNs with the topological information of the underlying graphs inherent in the graph spectra. To demonstrate this, we characterize the asymptotic behavior of GCNs on the Erd\H{o}s -- R\'{e}nyi graph. We show that when the Erd\H{o}s -- R\'{e}nyi graph is sufficiently dense and large, a broad range of GCNs on it suffers from the ``information loss" in the limit of infinite layers with high probability. Based on the theory, we provide a principled guideline for weight normalization of graph NNs. We experimentally confirm that the proposed weight scaling enhances the predictive performance of GCNs in real data. Code is available at https://github.com/delta2323/gnn-asymptotics.

Graph inference learning for semi-supervised classification

- Chunyan Xu, Zhen Cui, Xiaobin Hong, Tong Zhang, Jian Yang, Wei Liu
- abstract@open-review(Poster): In this work, we address the semi-supervised classification of graph data, where the categories of those unlabeled nodes are inferred from labeled nodes as well as graph structures. Recent works often solve this problem with the advanced graph convolution in a conventional supervised manner, but the performance could be heavily affected when labeled data is scarce. Here we propose a Graph Inference Learning (GIL) framework to boost the performance of node classification by learning the inference of node labels on graph topology. To bridge the connection of two nodes, we formally define a structure relation by encapsulating node attributes, between-node paths and local topological structures together, which can make inference conveniently deduced from one node to another node. For learning the inference process, we further introduce meta-optimization on structure relations from training nodes to validation nodes, such that the learnt graph inference capability can be better self-adapted into test nodes. Comprehensive evaluations on four benchmark datasets (including Cora, Citeseer, Pubmed and NELL) demonstrate the superiority of our GIL when compared with other state-of-the-art methods in the semi-supervised node classification task.

Sparse Coding with Gated Learned ISTA

- Kailun Wu, Yiwen Guo, Ziang Li, Changshui Zhang
- abstract@open-review(Spotlight): In this paper, we study the learned iterative shrinkage thresholding algorithm (LISTA) for solving sparse coding problems. Following assumptions made by prior works, we first discover that the code components in its estimations may be lower than expected, i.e., require gains, and to address this problem, a gated mechanism amenable to theoretical analysis is then introduced. Specific design of the gates is inspired by convergence analyses of the mechanism and hence its effectiveness can be formally guaranteed. In addition to the gain gates, we further introduce overshoot gates for compensating insufficient step size in LISTA. Extensive empirical results confirm our theoretical findings and verify the effectiveness of our method.

Learning deep graph matching with channel-independent embedding and Hungarian attention

- Tianshu Yu, Runzhong Wang, Junchi Yan, Baoxin Li
- abstract@open-review(Poster): Graph matching aims to establishing node-wise correspondence between two graphs, which is a classic combinatorial problem and in general NP-complete. Until very recently, deep graph matching methods start to resort to deep networks to achieve unprecedented matching accuracy. Along this direction, this paper makes two complementary contributions which can also be reused as plugin in existing works: i) a novel node and edge embedding strategy which stimulates the multi-head strategy in attention models and allows the information in each channel to be merged independently. In contrast, only node embedding is accounted in previous works; ii) a general masking mechanism over the loss function is devised to improve the smoothness of objective learning for graph matching. Using Hungarian algorithm, it dynamically constructs a structured and sparsely connected layer, taking into account the most contributing matching pairs as hard attention. Our approach performs competitively, and can also improve state-of-the-art methods as plugin, regarding with matching accuracy on three public benchmarks.

StructPool: Structured Graph Pooling via Conditional Random Fields

- Hao Yuan, Shuiwang Ji
- abstract@open-review(Poster): Learning high-level representations for graphs is of great importance for graph analysis tasks. In addition to graph convolution, graph pooling is an important but less explored research area. In particular, most of existing graph pooling techniques do not consider the graph structural information explicitly. We argue that such information is important and develop a novel graph pooling technique, know as the StructPool, in this work. We consider the graph pooling as a node clustering problem, which requires the learning of a cluster assignment matrix. We propose to formulate it as a structured prediction problem and employ conditional random fields to capture the relationships among assignments of different nodes. We also generalize our method to incorporate graph topological information in designing the Gibbs energy function. Experimental results on multiple datasets demonstrate the effectiveness of our proposed StructPool.

On the Weaknesses of Reinforcement Learning for Neural Machine Translation

- Leshem Choshen, Lior Fox, Zohar Aizenbud, Omri Abend
- abstract@open-review(Poster): Reinforcement learning (RL) is frequently used to increase performance in text generation tasks, including machine translation (MT), notably through the use of Minimum Risk Training (MRT) and Generative Adversarial Networks (GAN). However, little is known about what and how these methods learn in the context of MT. We prove that one of the most common RL methods for MT does not optimize the expected reward, as well as show that other methods take an infeasibly long time to converge. In fact, our results suggest that RL practices in MT are likely to improve performance only where the pre-trained parameters are already close to yielding the correct translation. Our findings further suggest that observed gains may be due to effects unrelated to the training signal, concretely, changes in the shape of the distribution curve.

Sharing Knowledge in Multi-Task Deep Reinforcement Learning

- Carlo D'Eramo, Davide Tateo, Andrea Bonarini, Marcello Restelli, Jan Peters
- abstract@open-review(Poster): We study the benefit of sharing representations among tasks to enable the effective use of deep neural networks in Multi-Task Reinforcement Learning. We leverage the assumption that learning from different tasks, sharing common properties, is helpful to generalize the knowledge of them resulting in a more effective feature extraction compared to learning a single task. Intuitively, the resulting set of features offers performance benefits when used by Reinforcement Learning algorithms. We prove this by providing theoretical guarantees that highlight the conditions for which is convenient to share representations among tasks, extending the well-known finite-time bounds of Approximate Value-Iteration to the multi-task setting. In addition, we complement our analysis by proposing multi-task extensions of three Reinforcement Learning algorithms that we empirically evaluate on widely used Reinforcement Learning benchmarks showing significant improvements over the single-task counterparts in terms of sample efficiency and performance.

Reinforcement Learning Based Graph-to-Sequence Model for Natural Question Generation

- Yu Chen, Lingfei Wu, Mohammed J. Zaki
- abstract@open-review(Poster): Natural question generation (QG) aims to generate questions from a passage and an answer. Previous works on QG either (i) ignore the rich structure information hidden in text, (ii) solely rely on cross-entropy loss that leads to issues like exposure bias and inconsistency between train/test measurement, or (iii) fail to fully exploit the answer information. To address these limitations, in this paper, we propose a reinforcement learning (RL) based graph-to-sequence (Graph2Seq) model for QG. Our model consists of a Graph2Seq generator with a novel Bidirectional Gated Graph Neural Network based encoder to embed the passage, and a hybrid evaluator with a mixed objective combining both cross-entropy and RL losses to ensure the generation of syntactically and semantically valid text. We also introduce an effective Deep Alignment Network for incorporating the answer information into the passage at both the word and contextual levels. Our model is end-to-end trainable and achieves new state-of-the-art scores, outperforming existing methods by a significant margin on the standard SQuAD benchmark.

SELF: Learning to Filter Noisy Labels with Self-Ensembling

- Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, Thomas Brox
- abstract@open-review(Poster): Deep neural networks (DNNs) have been shown to over-fit a dataset when being trained with noisy labels for a long enough time. To overcome this problem, we present a simple and effective method self-ensemble label filtering (SELF) to progressively filter out the wrong labels during training. Our method improves the task performance by gradually allowing supervision only from the potentially non-noisy (clean) labels and stops learning on the filtered noisy labels. For the filtering, we form running averages of predictions over the entire training dataset using the network output at different training epochs. We show that these ensemble estimates yield more accurate identification of inconsistent predictions throughout training than the single estimates of the network at the most recent training epoch. While filtered samples are removed entirely from the supervised training loss, we dynamically leverage them via semi-supervised learning in the unsupervised loss. We demonstrate the positive effect of such an approach on various image classification tasks under both symmetric and asymmetric label noise and at different noise ratios. It substantially outperforms all previous works on noise-aware learning across different datasets and can be applied to a broad set of network architectures.

Program Guided Agent

- Shao-Hua Sun, Te-Lin Wu, Joseph J. Lim
- abstract@open-review(Spotlight): Developing agents that can learn to follow natural language instructions has been an emerging research direction. While being accessible and flexible, natural language instructions can sometimes be ambiguous even to humans. To address this, we propose to utilize programs, structured in a formal language, as a precise and expressive way to specify tasks. We then devise a modular framework that learns to perform a task specified by a program – as different circumstances give rise to diverse ways to accomplish the task, our framework can perceive which circumstance it is currently under, and instruct a multitask policy accordingly to fulfill each subtask of the overall task. Experimental results on a 2D Minecraft environment not only demonstrate that the proposed framework learns to reliably accomplish program instructions and achieves zero-shot generalization to more complex instructions but also verify the efficiency of the proposed modulation mechanism for learning the multitask policy. We also conduct an analysis comparing various models which learn from programs and natural language instructions in an end-to-end fashion.

Large Batch Optimization for Deep Learning: Training BERT in 76 minutes

- Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, Cho-Jui Hsieh
- abstract@open-review(Poster): Training large deep neural networks on massive datasets is computationally very challenging. There has been recent surge in interest in using large batch stochastic optimization methods to tackle this issue. The most prominent algorithm in this line of research is LARS, which by employing layerwise adaptive learning rates trains ResNet on ImageNet in a few minutes. However, LARS performs poorly for attention models like BERT, indicating that its performance gains are not consistent across tasks. In this paper, we first study a principled layerwise adaptation strategy to accelerate training of deep neural networks using large mini-batches. Using this strategy, we develop a new layerwise adaptive large batch optimization technique called LAMB; we then provide convergence analysis of LAMB as well as LARS, showing convergence to a stationary point in general nonconvex settings. Our empirical results demonstrate the superior performance of LAMB across various tasks such as BERT and ResNet-50 training with very little hyperparameter tuning. In particular, for BERT training, our optimizer enables use of very large batch sizes of 32868 without any degradation of performance. By increasing the batch size to the memory limit of a TPUv3 Pod, BERT training time can be reduced from 3 days to just 76 minutes.