## Regularized Learning for Domain Adaptation under Label Shifts

- Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, Animashree Anandkumar
- abstract@open-review(Poster): We propose Regularized Learning under Label shifts (RLLS), a principled and a practical domain-adaptation algorithm to correct for shifts in the label distribution between a source and a target domain. We first estimate importance weights using labeled source data and unlabeled target data, and then train a classifier on the weighted source samples. We derive a generalization bound for the classifier on the target domain which is independent of the (ambient) data dimensions, and instead only depends on the complexity of the function class. To the best of our knowledge, this is the first generalization bound for the label-shift problem where the labels in the target domain are not available. Based on this bound, we propose a regularized estimator for the small-sample regime which accounts for the uncertainty in the estimated weights. Experiments on the CIFAR-10 and MNIST datasets show that RLLS improves classification accuracy, especially in the low sample and large-shift regimes, compared to previous methods.

## Towards Robust, Locally Linear Deep Networks

- Guang-He Lee, David Alvarez-Melis, Tommi S. Jaakkola
- abstract@open-review(Poster): Deep networks realize complex mappings that are often understood by their locally linear behavior at or around points of interest. For example, we use the derivative of the mapping with respect to its inputs for sensitivity analysis, or to explain (obtain coordinate relevance for) a prediction. One key challenge is that such derivatives are themselves inherently unstable. In this paper, we propose a new learning problem to encourage deep networks to have stable derivatives over larger regions. While the problem is challenging in general, we focus on networks with piecewise linear activation functions. Our algorithm consists of an inference step that identifies a region around a point where linear approximation is provably stable, and an optimization step to expand such regions. We propose a novel relaxation to scale the algorithm to realistic models. We illustrate our method with residual and recurrent networks on image and sequence datasets.

## The Limitations of Adversarial Training and the Blind-Spot Attack

- Huan Zhang, *Hongge Chen*, Zhao Song, Duane Boning, Inderjit S. Dhillon, Cho-Jui Hsieh
- abstract@open-review(Poster): The adversarial training procedure proposed by Madry et al. (2018) is one of the most effective methods to defend against adversarial examples in deep neural net- works (DNNs). In our paper, we shed some lights on the practicality and the hardness of adversarial training by showing that the effectiveness (robustness on test set) of adversarial training has a strong correlation with the distance between a test point and the manifold of training data embedded by the network. Test examples that are relatively far away from this manifold are more likely to be vulnerable to adversarial attacks. Consequentially, an adversarial training based defense is susceptible to a new class of attacks, the “blind-spot attack”, where the input images reside in “blind-spots” (low density regions) of the empirical distri- bution of training data but is still on the ground-truth data manifold. For MNIST, we found that these blind-spots can be easily found by simply scaling and shifting image pixel values. Most importantly, for large datasets with high dimensional and complex data manifold (CIFAR, ImageNet, etc), the existence of blind-spots in adversarial training makes defending on any valid test examples difficult due to the curse of dimensionality and the scarcity of training data. Additionally, we find that blind-spots also exist on provable defenses including (Kolter & Wong, 2018) and (Sinha et al., 2018) because these trainable robustness certificates can only be practically optimized on a limited set of training data.

## Learning to Learn without Forgetting by Maximizing Transfer and Minimizing Interference

- Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro
- abstract@open-review(Poster): Lack of performance when it comes to continual learning over non-stationary distributions of data remains a major challenge in scaling neural network learning to more human realistic settings. In this work we propose a new conceptualization of the continual learning problem in terms of a temporally symmetric trade-off between transfer and interference that can be optimized by enforcing gradient alignment across examples. We then propose a new algorithm, Meta-Experience Replay (MER), that directly exploits this view by combining experience replay with optimization based meta-learning. This method learns parameters that make interference based on future gradients less likely and transfer based on future gradients more likely. We conduct experiments across continual lifelong supervised learning benchmarks and non-stationary reinforcement learning environments demonstrating that our approach consistently outperforms recently proposed baselines for continual learning. Our experiments show that the gap between the performance of MER and baseline algorithms grows both as the environment gets more non-stationary and as the fraction of the total experiences stored gets smaller.

## Global-to-local Memory Pointer Networks for Task-Oriented Dialogue

- Chien-Sheng Wu, Richard Socher, Caiming Xiong
- abstract@open-review(Poster): End-to-end task-oriented dialogue is challenging since knowledge bases are usually large, dynamic and hard to incorporate into a learning framework. We propose the global-to-local memory pointer (GLMP) networks to address this issue. In our model, a global memory encoder and a local memory decoder are proposed to share external knowledge. The encoder encodes dialogue history, modifies global contextual representation, and generates a global memory pointer. The decoder first generates a sketch response with unfilled slots. Next, it passes the global memory pointer to filter the external knowledge for relevant information, then instantiates the slots via the local memory pointers. We empirically show that our model can improve copy accuracy and mitigate the common out-of-vocabulary problem. As a result, GLMP is able to improve over the previous state-of-the-art models in both simulated bAbI Dialogue dataset and human-human Stanford Multi-domain Dialogue dataset on automatic and human evaluation.

## Rethinking the Value of Network Pruning

- Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, Trevor Darrell
- abstract@open-review(Poster): Network pruning is widely used for reducing the heavy inference cost of deep models in low-resource settings. A typical pruning algorithm is a three-stage pipeline, i.e., training (a large model), pruning and fine-tuning. During pruning, according to a certain criterion, redundant weights are pruned and important weights are kept to best preserve the accuracy. In this work, we make several surprising observations which contradict common beliefs. For all state-of-the-art structured pruning algorithms we examined, fine-tuning a pruned model only gives comparable or worse performance than training that model with randomly initialized weights. For pruning algorithms which assume a predefined target network architecture, one can get rid of the full pipeline and directly train the target network from scratch. Our observations are consistent for multiple network architectures, datasets, and tasks, which imply that: 1) training a large, over-parameterized model is often not necessary to obtain an efficient final model, 2) learned `important'' weights of the large model are typically not useful for the small pruned model, 3) the pruned architecture itself, rather than a set of inherited`important" weights, is more crucial to the efficiency in the final model, which suggests that in some cases pruning can be useful as an architecture search paradigm. Our results suggest the need for more careful baseline evaluations in future research on structured pruning methods. We also compare with the "Lottery Ticket Hypothesis" (Frankle & Carbin 2019), and find that with optimal learning rate, the "winning ticket" initialization as used in Frankle & Carbin (2019) does not bring improvement over random initialization.

## Neural TTS Stylization with Adversarial and Collaborative Games

- Shuang Ma, Daniel Mcduff, Yale Song

- abstract@open-review(Poster): The modeling of style when synthesizing natural human speech from text has been the focus of significant attention. Some state-of-the-art approaches train an encoder-decoder network on paired text and audio samples (x_txt, x_aud) by encouraging its output to reconstruct x_aud. The synthesized audio waveform is expected to contain the verbal content of x_txt and the auditory style of x_aud. Unfortunately, modeling style in TTS is somewhat under-determined and training models with a reconstruction loss alone is insufficient to disentangle content and style from other factors of variation. In this work, we introduce an end-to-end TTS model that offers enhanced content-style disentanglement ability and controllability. We achieve this by combining a pairwise training procedure, an adversarial game, and a collaborative game into one training scheme. The adversarial game concentrates the true data distribution, and the collaborative game minimizes the distance between real samples and generated samples in both the original space and the latent space. As a result, the proposed model delivers a highly controllable generator, and a disentangled representation. Benefiting from the separate modeling of style and content, our model can generate human fidelity speech that satisfies the desired style conditions. Our model achieves start-of-the-art results across multiple tasks, including style transfer (content and style swapping), emotion modeling, and identity transfer (fitting a new speaker's voice).

## On the Universal Approximability and Complexity Bounds of Quantized ReLU Neural Networks

- Yukun Ding, Jinglan Liu, Jinjun Xiong, Yiyu Shi
- abstract@open-review(Poster): Compression is a key step to deploy large neural networks on resource-constrained platforms. As a popular compression technique, quantization constrains the number of distinct weight values and thus reducing the number of bits required to represent and store each weight. In this paper, we study the representation power of quantized neural networks. First, we prove the universal approximability of quantized ReLU networks on a wide class of functions. Then we provide upper bounds on the number of weights and the memory size for a given approximation error bound and the bit-width of weights for function-independent and function-dependent structures. Our results reveal that, to attain an approximation error bound of $\epsilon$, the number of weights needed by a quantized network is no more than $\mathcal{O}\left(\log^5(1/\epsilon)\right)$ times that of an unquantized network. This overhead is of much lower order than the lower bound of the number of weights needed for the error bound, supporting the empirical success of various quantization techniques. To the best of our knowledge, this is the first in-depth study on the complexity bounds of quantized neural networks.

## Poincare Glove: Hyperbolic Word Embeddings

- Alexandru Tifrea, *Gary Becigneul*, Octavian-Eugen Ganea*
- abstract@open-review(Poster): Words are not created equal. In fact, they form an aristocratic graph with a latent hierarchical structure that the next generation of unsupervised learned word embeddings should reveal. In this paper, justified by the notion of delta-hyperbolicity or tree-likeliness of a space, we propose to embed words in a Cartesian product of hyperbolic spaces which we theoretically connect to the Gaussian word embeddings and their Fisher geometry. This connection allows us to introduce a novel principled hypernymy score for word embeddings. Moreover, we adapt the well-known Glove algorithm to learn unsupervised word embeddings in this type of Riemannian manifolds. We further explain how to solve the analogy task using the Riemannian parallel transport that generalizes vector arithmetics to this new type of geometry. Empirically, based on extensive experiments, we prove that our embeddings, trained unsupervised, are the first to simultaneously outperform strong and popular baselines on the tasks of similarity, analogy and hypernymy detection. In particular, for word hypernymy, we obtain new state-of-the-art on fully unsupervised WBLESS classification accuracy.

## Eidetic 3D LSTM: A Model for Video Prediction and Beyond

- Yunbo Wang, Lu Jiang, Ming-Hsuan Yang, Li-Jia Li, Mingsheng Long, Li Fei-Fei
- abstract@open-review(Poster): Spatiotemporal predictive learning, though long considered to be a promising self-supervised feature learning method, seldom shows its effectiveness beyond future video prediction. The reason is that it is difficult to learn good representations for both short-term frame dependency and long-term high-level relations. We present a new model, Eidetic 3D LSTM (E3D-LSTM), that integrates 3D convolutions into RNNs. The encapsulated 3D-Conv makes local perceptrons of RNNs motion-aware and enables the memory cell to store better short-term features. For long-term relations, we make the present memory state interact with its historical records via a gate-controlled self-attention module. We describe this memory transition mechanism eidetic as it is able to effectively recall the stored memories across multiple time stamps even after long periods of disturbance. We first evaluate the E3D-LSTM network on widely-used future video prediction datasets and achieve the state-of-the-art performance. Then we show that the E3D-LSTM network also performs well on the early activity recognition to infer what is happening or what will happen after observing only limited frames of video. This task aligns well with video prediction in modeling action intentions and tendency.

## Towards GAN Benchmarks Which Require Generalization

- Ishaan Gulrajani, Colin Raffel, Luke Metz
- abstract@open-review(Poster): For many evaluation metrics commonly used as benchmarks for unconditional image generation, trivially memorizing the training set attains a better score than models which are considered state-of-the-art; we consider this problematic. We clarify a necessary condition for an evaluation metric not to behave this way: estimating the function must require a large sample from the model. In search of such a metric, we turn to neural network divergences (NNDs), which are defined in terms of a neural network trained to distinguish between distributions. The resulting benchmarks cannot be ``won'' by training set memorization, while still being perceptually correlated and computable only from samples. We survey past work on using NNDs for evaluation, implement an example black-box metric based on these ideas, and validate experimentally that it can measure a notion of generalization.

## There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average

- Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, Andrew Gordon Wilson
- abstract@open-review(Poster): Presently the most successful approaches to semi-supervised learning are based on consistency regularization, whereby a model is trained to be robust to small perturbations of its inputs and parameters. To understand consistency regularization, we conceptually explore how loss geometry interacts with training procedures. The consistency loss dramatically improves generalization performance over supervised-only training; however, we show that SGD struggles to converge on the consistency loss and continues to make large steps that lead to changes in predictions on the test data. Motivated by these observations, we propose to train consistency-based methods with Stochastic Weight Averaging (SWA), a recent approach which averages weights along the trajectory of SGD with a modified learning rate schedule. We also propose fast-SWA, which further accelerates convergence by averaging multiple points within each cycle of a cyclical learning rate schedule. With weight averaging, we achieve the best known semi-supervised results on CIFAR-10 and CIFAR-100, over many different quantities of labeled training data. For example, we achieve 5.0% error on CIFAR-10 with only 4000 labels, compared to the previous best result in the literature of 6.3%.

## Synthetic Datasets for Neural Program Synthesis

- Richard Shin, Neel Kant, Kavi Gupta, Chris Bender, Brandon Trabucco, Rishabh Singh, Dawn Song
- abstract@open-review(Poster): The goal of program synthesis is to automatically generate programs in a particular language from corresponding specifications, e.g. input-output behavior. Many current approaches achieve impressive results after training on randomly generated I/O examples in limited domain-specific languages (DSLs), as with string transformations in RobustFill. However, we empirically discover that applying test input generation techniques for languages with control flow and rich input space causes deep networks to generalize poorly to certain data distributions; to correct this, we propose a new methodology for controlling and evaluating the bias of synthetic data distributions over both programs and specifications.

We demonstrate, using the Karel DSL and a small Calculator DSL, that training deep networks on these distributions leads to improved cross-distribution generalization performance.

## Stochastic Prediction of Multi-Agent Interactions from Partial Observations

- Chen Sun, Per Karlsson, Jiajun Wu, Joshua B Tenenbaum, Kevin Murphy
- abstract@open-review(Poster): We present a method which learns to integrate temporal information, from a learned dynamics model, with ambiguous visual information, from a learned vision model, in the context of interacting agents. Our method is based on a graph-structured variational recurrent neural network, which is trained end-to-end to infer the current state of the (partially observed) world, as well as to forecast future states. We show that our method outperforms various baselines on two sports datasets, one based on real basketball trajectories, and one generated by a soccer game engine.

## DyRep: Learning Representations over Dynamic Graphs

- Rakshit Trivedi, Mehrdad Farajtabar, Prasenjeet Biswal, Hongyuan Zha
- abstract@open-review(Poster): Representation Learning over graph structured data has received significant attention recently due to its ubiquitous applicability. However, most advancements have been made in static graph settings while efforts for jointly learning dynamic of the graph and dynamic on the graph are still in an infant stage. Two fundamental questions arise in learning over dynamic graphs: (i) How to elegantly model dynamical processes over graphs? (ii) How to leverage such a model to effectively encode evolving graph information into low-dimensional representations? We present DyRep - a novel modeling framework for dynamic graphs that posits representation learning as a latent mediation process bridging two observed processes namely -- dynamics of the network (realized as topological evolution) and dynamics on the network (realized as activities between nodes). Concretely, we propose a two-time scale deep temporal point process model that captures the interleaved dynamics of the observed processes. This model is further parameterized by a temporal-attentive representation network that encodes temporally evolving structural information into node representations which in turn drives the nonlinear evolution of the observed graph dynamics. Our unified framework is trained using an efficient unsupervised procedure and has capability to generalize over unseen nodes. We demonstrate that DyRep outperforms state-of-the-art baselines for dynamic link prediction and time prediction tasks and present extensive qualitative insights into our framework.

## Label super-resolution networks

- Kolya Malkin, Caleb Robinson, Le Hou, Rachel Soobitsky, Jacob Czawlytko, Dimitris Samaras, Joel Saltz, Lucas Joppa, Nebojsa Jojic
- abstract@open-review(Poster): We present a deep learning-based method for super-resolving coarse (low-resolution) labels assigned to groups of image pixels into pixel-level (high-resolution) labels, given the joint distribution between those low- and high-resolution labels. This method involves a novel loss function that minimizes the distance between a distribution determined by a set of model outputs and the corresponding distribution given by low-resolution labels over the same set of outputs. This setup does not require that the high-resolution classes match the low-resolution classes and can be used in high-resolution semantic segmentation tasks where high-resolution labeled data is not available. Furthermore, our proposed method is able to utilize both data with low-resolution labels and any available high-resolution labels, which we show improves performance compared to a network trained only with the same amount of high-resolution data. We test our proposed algorithm in a challenging land cover mapping task to super-resolve labels at a 30m resolution to a separate set of labels at a 1m resolution. We compare our algorithm with models that are trained on high-resolution data and show that 1) we can achieve similar performance using only low-resolution data; and 2) we can achieve better performance when we incorporate a small amount of high-resolution data in our training. We also test our approach on a medical imaging problem, resolving low-resolution probability maps into high-resolution segmentation of lymphocytes with accuracy equal to that of fully supervised models.

## Multi-step Retriever-Reader Interaction for Scalable Open-domain Question Answering

- Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, Andrew McCallum
- abstract@open-review(Poster): This paper introduces a new framework for open-domain question answering in which the retriever and the reader \emph{iteratively interact} with each other. The framework is agnostic to the architecture of the machine reading model provided it has \emph{access} to the token-level hidden representations of the reader. The retriever uses fast nearest neighbor search that allows it to scale to corpora containing millions of paragraphs. A gated recurrent unit updates the query at each step conditioned on the \emph{state} of the reader and the \emph{reformulated} query is used to re-rank the paragraphs by the retriever. We conduct analysis and show that iterative interaction helps in retrieving informative paragraphs from the corpus. Finally, we show that our multi-step-reasoning framework brings consistent improvement when applied to two widely used reader architectures (\drqa and \bidaf) on various large open-domain datasets ---\tqau, \quasart, \searchqa, and \squado\footnote{Code and pretrained models are available at \url{https://github.com/rajarshd/Multi-Step-Reasoning}}.

## Capsule Graph Neural Network

- Zhang Xinyi, Lihui Chen
- abstract@open-review(Poster): The high-quality node embeddings learned from the Graph Neural Networks (GNNs) have been applied to a wide range of node-based applications and some of them have achieved state-of-the-art (SOTA) performance. However, when applying node embeddings learned from GNNs to generate graph embeddings, the scalar node representation may not suffice to preserve the node/graph properties efficiently, resulting in sub-optimal graph embeddings.

Inspired by the Capsule Neural Network (CapsNet), we propose the Capsule Graph Neural Network (CapsGNN), which adopts the concept of capsules to address the weakness in existing GNN-based graph embeddings algorithms. By extracting node features in the form of capsules, routing mechanism can be utilized to capture important information at the graph level. As a result, our model generates multiple embeddings for each graph to capture graph properties from different aspects. The attention module incorporated in CapsGNN is used to tackle graphs with various sizes which also enables the model to focus on critical parts of the graphs.

Our extensive evaluations with 10 graph-structured datasets demonstrate that CapsGNN has a powerful mechanism that operates to capture macroscopic properties of the whole graph by data-driven. It outperforms other SOTA techniques on several graph classification tasks, by virtue of the new instrument.

## Learning a Meta-Solver for Syntax-Guided Program Synthesis

- Xujie Si, Yuan Yang, Hanjun Dai, Mayur Naik, Le Song
- abstract@open-review(Poster): We study a general formulation of program synthesis called syntax-guided synthesis(SyGuS) that concerns synthesizing a program that follows a given grammar and satisfies a given logical specification. Both the logical specification and the grammar have complex structures and can vary from task to task, posing significant challenges for learning across different tasks. Furthermore, training data is often unavailable for domain specific synthesis tasks. To address these challenges, we propose a meta-learning framework that learns a transferable policy from only weak supervision. Our framework consists of three components: 1) an encoder, which embeds both the logical specification and grammar at the same time using a graph neural network; 2) a grammar adaptive policy network which enables learning a transferable policy; and 3) a reinforcement learning algorithm that jointly trains the embedding and adaptive policy. We evaluate the framework on 214 cryptographic circuit synthesis tasks. It solves 141 of them in the out-of-box solver setting, significantly outperforming a similar search-based approach but without learning, which solves only 31. The result is comparable to two

state-of-the-art classical synthesis engines, which solve 129 and 153 respectively. In the meta-solver setting, the framework can efficiently adapt to unseen tasks and achieves speedup ranging from 2x up to 100x.

## Stochastic Optimization of Sorting Networks via Continuous Relaxations

- Aditya Grover, Eric Wang, Aaron Zweig, Stefano Ermon
- abstract@open-review(Poster): Sorting input objects is an important step in many machine learning pipelines. However, the sorting operator is non-differentiable with respect to its inputs, which prohibits end-to-end gradient-based optimization. In this work, we propose NeuralSort, a general-purpose continuous relaxation of the output of the sorting operator from permutation matrices to the set of unimodal row-stochastic matrices, where every row sums to one and has a distinct argmax. This relaxation permits straight-through optimization of any computational graph involve a sorting operation. Further, we use this relaxation to enable gradient-based stochastic optimization over the combinatorially large space of permutations by deriving a reparameterized gradient estimator for the Plackett-Luce family of distributions over permutations. We demonstrate the usefulness of our framework on three tasks that require learning semantic orderings of high-dimensional objects, including a fully differentiable, parameterized extension of the k-nearest neighbors algorithm

## Energy-Constrained Compression for Deep Neural Networks via Weighted Sparse Projection and Layer Input Masking

- Haichuan Yang, Yuhao Zhu, Ji Liu
- abstract@open-review(Poster): Deep Neural Networks (DNNs) are increasingly deployed in highly energy-constrained environments such as autonomous drones and wearable devices while at the same time must operate in real-time. Therefore, reducing the energy consumption has become a major design consideration in DNN training. This paper proposes the first end-to-end DNN training framework that provides quantitative energy consumption guarantees via weighted sparse projection and input masking. The key idea is to formulate the DNN training as an optimization problem in which the energy budget imposes a previously unconsidered optimization constraint. We integrate the quantitative DNN energy estimation into the DNN training process to assist the constrained optimization. We prove that an approximate algorithm can be used to efficiently solve the optimization problem. Compared to the best prior energy-saving techniques, our framework trains DNNs that provide higher accuracies under same or lower energy budgets.

## Composing Complex Skills by Learning Transition Policies

- Youngwoon Lee, *Shao-Hua Sun*, Sriram Somasundaram, Edward S. Hu, Joseph J. Lim
- abstract@open-review(Poster): Humans acquire complex skills by exploiting previously learned skills and making transitions between them. To empower machines with this ability, we propose a method that can learn transition policies which effectively connect primitive skills to perform sequential tasks without handcrafted rewards. To efficiently train our transition policies, we introduce proximity predictors which induce rewards gauging proximity to suitable initial states for the next skill. The proposed method is evaluated on a set of complex continuous control tasks in bipedal locomotion and robotic arm manipulation which traditional policy gradient methods struggle at. We demonstrate that transition policies enable us to effectively compose complex skills with existing primitive skills. The proposed induced rewards computed using the proximity predictor further improve training efficiency by providing more dense information than the sparse rewards from the environments. We make our environments, primitive skills, and code public for further research at https://youngwoon.github.io/transition .

## Revealing interpretable object representations from human behavior

- Charles Y. Zheng, Francisco Pereira, Chris I. Baker, Martin N. Hebart
- abstract@open-review(Poster): To study how mental object representations are related to behavior, we estimated sparse, non-negative representations of objects using human behavioral judgments on images representative of 1,854 object categories. These representations predicted a latent similarity structure between objects, which captured most of the explainable variance in human behavioral judgments. Individual dimensions in the low-dimensional embedding were found to be highly reproducible and interpretable as conveying degrees of taxonomic membership, functionality, and perceptual attributes. We further demonstrated the predictive power of the embeddings for explaining other forms of human behavior, including categorization, typicality judgments, and feature ratings, suggesting that the dimensions reflect human conceptual representations of objects beyond the specific task.

## ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware

- Han Cai, Ligeng Zhu, Song Han
- abstract@open-review(Poster): Neural architecture search (NAS) has a great impact by automatically designing effective neural network architectures. However, the prohibitive computational demand of conventional NAS algorithms (e.g. 10 4 GPU hours) makes it difficult to directly search the architectures on large-scale tasks (e.g. ImageNet). Differentiable NAS can reduce the cost of GPU hours via a continuous representation of network architecture but suffers from the high GPU memory consumption issue (grow linearly w.r.t. candidate set size). As a result, they need to utilize proxy tasks, such as training on a smaller dataset, or learning with only a few blocks, or training just for a few epochs. These architectures optimized on proxy tasks are not guaranteed to be optimal on the target task. In this paper, we present ProxylessNAS that can directly learn the architectures for large-scale target tasks and target hardware platforms. We address the high memory consumption issue of differentiable NAS and reduce the computational cost (GPU hours and GPU memory) to the same level of regular training while still allowing a large candidate set. Experiments on CIFAR-10 and ImageNet demonstrate the effectiveness of directness and specialization. On CIFAR-10, our model achieves 2.08% test error with only 5.7M parameters, better than the previous state-of-the-art architecture AmoebaNet-B, while using 6× fewer parameters. On ImageNet, our model achieves 3.1% better top-1 accuracy than MobileNetV2, while being 1.2× faster with measured GPU latency. We also apply ProxylessNAS to specialize neural architectures for hardware with direct hardware metrics (e.g. latency) and provide insights for efficient CNN architecture design.

## A Generative Model For Electron Paths

- John Bradshaw, Matt J. Kusner, Brooks Paige, Marwin H. S. Segler, José Miguel Hernández-Lobato
- abstract@open-review(Poster): Chemical reactions can be described as the stepwise redistribution of electrons in molecules. As such, reactions are often depicted using "arrow-pushing" diagrams which show this movement as a sequence of arrows. We propose an electron path prediction model (ELECTRO) to learn these sequences directly from raw reaction data. Instead of predicting product molecules directly from reactant molecules in one shot, learning a model of electron movement has the benefits of (a) being easy for chemists to interpret, (b) incorporating constraints of chemistry, such as balanced atom counts before and after the reaction, and (c) naturally encoding the sparsity of chemical reactions, which usually involve changes in only a small number of atoms in the reactants. We design a method to extract approximate reaction paths from any dataset of atom-mapped reaction SMILES strings. Our model achieves excellent performance on an important subset of the USPTO reaction dataset, comparing favorably to the strongest baselines. Furthermore, we show that our model recovers a basic knowledge of chemistry without being explicitly trained to do so.

## Learning to Infer and Execute 3D Shape Programs

- Yonglong Tian, Andrew Luo, Xingyuan Sun, Kevin Ellis, William T. Freeman, Joshua B. Tenenbaum, Jiajun Wu
- abstract@open-review(Poster): Human perception of 3D shapes goes beyond reconstructing them as a set of points or a composition of geometric primitives: we also effortlessly understand higher-level shape structure such as the repetition and reflective symmetry of object parts. In contrast, recent

advances in 3D shape sensing focus more on low-level geometry but less on these higher-level relationships. In this paper, we propose 3D shape programs, integrating bottom-up recognition systems with top-down, symbolic program structure to capture both low-level geometry and high-level structural priors for 3D shapes. Because there are no annotations of shape programs for real shapes, we develop neural modules that not only learn to infer 3D shape programs from raw, unannotated shapes, but also to execute these programs for shape reconstruction. After initial bootstrapping, our end-to-end differentiable model learns 3D shape programs by reconstructing shapes in a self-supervised manner. Experiments demonstrate that our model accurately infers and executes 3D shape programs for highly complex shapes from various categories. It can also be integrated with an image-to-shape module to infer 3D shape programs directly from an RGB image, leading to 3D shape reconstructions that are both more accurate and more physically plausible.

## Music Transformer: Generating Music with Long-Term Structure

- Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, Douglas Eck
- abstract@open-review(Poster): Music relies heavily on repetition to build structure and meaning. Self-reference occurs on multiple timescales, from motifs to phrases to reusing of entire sections of music, such as in pieces with ABA structure. The Transformer (Vaswani et al., 2017), a sequence model based on self-attention, has achieved compelling results in many generation tasks that require maintaining long-range coherence. This suggests that self-attention might also be well-suited to modeling music. In musical composition and performance, however, relative timing is critically important. Existing approaches for representing relative positional information in the Transformer modulate attention based on pairwise distance (Shaw et al., 2018). This is impractical for long sequences such as musical compositions since their memory complexity is quadratic in the sequence length. We propose an algorithm that reduces the intermediate memory requirements to linear in the sequence length. This enables us to demonstrate that a Transformer with our modified relative attention mechanism can generate minute-long (thousands of steps) compositions with compelling structure, generate continuations that coherently elaborate on a given motif, and in a seq2seq setup generate accompaniments conditioned on melodies. We evaluate the Transformer with our relative attention mechanism on two datasets, JSB Chorales and Piano-e-competition, and obtain state-of-the-art results on the latter.

## Modeling the Long Term Future in Model-Based Reinforcement Learning

- Nan Rosemary Ke, Amanpreet Singh, Ahmed Touati, Anirudh Goyal, Yoshua Bengio, Devi Parikh, Dhruv Batra
- abstract@open-review(Poster): In model-based reinforcement learning, the agent interleaves between model learning and planning. These two components are inextricably intertwined. If the model is not able to provide sensible long-term prediction, the executed planer would exploit model flaws, which can yield catastrophic failures. This paper focuses on building a model that reasons about the long-term future and demonstrates how to use this for efficient planning and exploration. To this end, we build a latent-variable autoregressive model by leveraging recent ideas in variational inference. We argue that forcing latent variables to carry future information through an auxiliary task substantially improves long-term predictions. Moreover, by planning in the latent space, the planner's solution is ensured to be within regions where the model is valid. An exploration strategy can be devised by searching for unlikely trajectories under the model. Our methods achieves higher reward faster compared to baselines on a variety of tasks and environments in both the imitation learning and model-based reinforcement learning settings.

## Model-Predictive Policy Learning with Uncertainty Regularization for Driving in Dense Traffic

- Mikael Henaff, Alfredo Canziani, Yann LeCun
- abstract@open-review(Poster): Learning a policy using only observational data is challenging because the distribution of states it induces at execution time may differ from the distribution observed during training. In this work, we propose to train a policy while explicitly penalizing the mismatch between these two distributions over a fixed time horizon. We do this by using a learned model of the environment dynamics which is unrolled for multiple time steps, and training a policy network to minimize a differentiable cost over this rolled-out trajectory. This cost contains two terms: a policy cost which represents the objective the policy seeks to optimize, and an uncertainty cost which represents its divergence from the states it is trained on. We propose to measure this second cost by using the uncertainty of the dynamics model about its own predictions, using recent ideas from uncertainty estimation for deep networks. We evaluate our approach using a large-scale observational dataset of driving behavior recorded from traffic cameras, and show that we are able to learn effective driving policies from purely observational data, with no environment interaction.

## Learning to Screen for Fast Softmax Inference on Large Vocabulary Neural Networks

- Patrick Chen, Si Si, Sanjiv Kumar, Yang Li, Cho-Jui Hsieh
- abstract@open-review(Poster): Neural language models have been widely used in various NLP tasks, including machine translation, next word prediction and conversational agents. However, it is challenging to deploy these models on mobile devices due to their slow prediction speed, where the bottleneck is to compute top candidates in the softmax layer. In this paper, we introduce a novel softmax layer approximation algorithm by exploiting the clustering structure of context vectors. Our algorithm uses a light-weight screening model to predict a much smaller set of candidate words based on the given context, and then conducts an exact softmax only within that subset. Training such a procedure end-to-end is challenging as traditional clustering methods are discrete and non-differentiable, and thus unable to be used with back-propagation in the training process. Using the Gumbel softmax, we are able to train the screening model end-to-end on the training set to exploit data distribution. The algorithm achieves an order of magnitude faster inference than the original softmax layer for predicting top-k words in various tasks such as beam search in machine translation or next words prediction. For example, for machine translation task on German to English dataset with around 25K vocabulary, we can achieve 20.4 times speed up with 98.9% precision@1 and 99.3% precision@5 with the original softmax layer prediction, while state-of-the-art (Zhang et al., 2018) only achieves 6.7x speedup with 98.7% precision@1 and 98.1% precision@5 for the same task.

## Interpolation-Prediction Networks for Irregularly Sampled Time Series

- Satya Narayan Shukla, Benjamin Marlin
- abstract@open-review(Poster): In this paper, we present a new deep learning architecture for addressing the problem of supervised learning with sparse and irregularly sampled multivariate time series. The architecture is based on the use of a semi-parametric interpolation network followed by the application of a prediction network. The interpolation network allows for information to be shared across multiple dimensions of a multivariate time series during the interpolation stage, while any standard deep learning model can be used for the prediction network. This work is motivated by the analysis of physiological time series data in electronic health records, which are sparse, irregularly sampled, and multivariate. We investigate the performance of this architecture on both classification and regression tasks, showing that our approach outperforms a range of baseline and recently proposed models.

## Contingency-Aware Exploration in Reinforcement Learning

- Jongwook Choi, Yijie Guo, Marcin Moczulski, Junhyuk Oh, Neal Wu, Mohammad Norouzi, Honglak Lee
- abstract@open-review(Poster): This paper investigates whether learning contingency-awareness and controllable aspects of an environment can lead to better exploration in reinforcement learning. To investigate this question, we consider an instantiation of this hypothesis evaluated on the Arcade Learning Element (ALE). In this study, we develop an attentive dynamics model (ADM) that discovers controllable elements of the observations, which are often associated with the location of the character in Atari games. The ADM is trained in a self-supervised fashion to predict the actions taken by the agent. The learned contingency information is used as a part of the state representation for exploration purposes. We demonstrate that combining actor-critic algorithm with count-based exploration using our representation achieves impressive results on a set of notoriously challenging Atari games due to sparse rewards. For example, we report a state-of-the-art score of >11,000 points on Montezuma's Revenge without using expert demonstrations, explicit high-

level information (e.g., RAM states), or supervisory data. Our experiments confirm that contingency-awareness is indeed an extremely powerful concept for tackling exploration problems in reinforcement learning and opens up interesting research questions for further investigations.

## Neural Graph Evolution: Towards Efficient Automatic Robot Design

- Tingwu Wang, Yuhao Zhou, Sanja Fidler, Jimmy Ba
- abstract@open-review(Poster): Despite the recent successes in robotic locomotion control, the design of robot relies heavily on human engineering. Automatic robot design has been a long studied subject, but the recent progress has been slowed due to the large combinatorial search space and the difficulty in evaluating the found candidates. To address the two challenges, we formulate automatic robot design as a graph search problem and perform evolution search in graph space. We propose Neural Graph Evolution (NGE), which performs selection on current candidates and evolves new ones iteratively. Different from previous approaches, NGE uses graph neural networks to parameterize the control policies, which reduces evaluation cost on new candidates with the help of skill transfer from previously evaluated designs. In addition, NGE applies Graph Mutation with Uncertainty (GM-UC) by incorporating model uncertainty, which reduces the search space by balancing exploration and exploitation. We show that NGE significantly outperforms previous methods by an order of magnitude. As shown in experiments, NGE is the first algorithm that can automatically discover kinematically preferred robotic graph structures, such as a fish with two symmetrical flat side-fins and a tail, or a cheetah with athletic front and back legs. Instead of using thousands of cores for weeks, NGE efficiently solves searching problem within a day on a single 64 CPU-core Amazon EC2 machine.

## Selfless Sequential Learning

- Rahaf Aljundi, Marcus Rohrbach, Tinne Tuytelaars
- abstract@open-review(Poster): Sequential learning, also called lifelong learning, studies the problem of learning tasks in a sequence with access restricted to only the data of the current task. In this paper we look at a scenario with fixed model capacity, and postulate that the learning process should not be selfish, i.e. it should account for future tasks to be added and thus leave enough capacity for them. To achieve Selfless Sequential Learning we study different regularization strategies and activation functions. We find that imposing sparsity at the level of the representation (i.e. neuron activations) is more beneficial for sequential learning than encouraging parameter sparsity. In particular, we propose a novel regularizer, that encourages representation sparsity by means of neural inhibition. It results in few active neurons which in turn leaves more free neurons to be utilized by upcoming tasks. As neural inhibition over an entire layer can be too drastic, especially for complex tasks requiring strong representations, our regularizer only inhibits other neurons in a local neighbourhood, inspired by lateral inhibition processes in the brain. We combine our novel regularizer with state-of-the-art lifelong learning methods that penalize changes to important previously learned parts of the network. We show that our new regularizer leads to increased sparsity which translates in consistent performance improvement on diverse datasets.

## Learning Particle Dynamics for Manipulating Rigid Bodies, Deformable Objects, and Fluids

- Yunzhu Li, Jiajun Wu, Russ Tedrake, Joshua B. Tenenbaum, Antonio Torralba
- abstract@open-review(Poster): Real-life control tasks involve matters of various substances---rigid or soft bodies, liquid, gas---each with distinct physical behaviors. This poses challenges to traditional rigid-body physics engines. Particle-based simulators have been developed to model the dynamics of these complex scenes; however, relying on approximation techniques, their simulation often deviates from real-world physics, especially in the long term. In this paper, we propose to learn a particle-based simulator for complex control tasks. Combining learning with particle-based systems brings in two major benefits: first, the learned simulator, just like other particle-based systems, acts widely on objects of different materials; second, the particle-based representation poses strong inductive bias for learning: particles of the same type have the same dynamics within. This enables the model to quickly adapt to new environments of unknown dynamics within a few observations. We demonstrate robots achieving complex manipulation tasks using the learned simulator, such as manipulating fluids and deformable foam, with experiments both in simulation and in the real world. Our study helps lay the foundation for robot learning of dynamic scenes with particle-based representations.

## Representing Formal Languages: A Comparison Between Finite Automata and Recurrent Neural Networks

- Joshua J. Michalenko, Ameesh Shah, Abhinav Verma, Richard G. Baraniuk, Swarat Chaudhuri, Ankit B. Patel
- abstract@open-review(Poster): We investigate the internal representations that a recurrent neural network (RNN) uses while learning to recognize a regular formal language. Specifically, we train a RNN on positive and negative examples from a regular language, and ask if there is a simple decoding function that maps states of this RNN to states of the minimal deterministic finite automaton (MDFA) for the language. Our experiments show that such a decoding function indeed exists, and that it maps states of the RNN not to MDFA states, but to states of an {\em abstraction} obtained by clustering small sets of MDFA states into ``"superstates". A qualitative analysis reveals that the abstraction often has a simple interpretation. Overall, the results suggest a strong structural relationship between internal representations used by RNNs and finite automata, and explain the well-known ability of RNNs to recognize formal grammatical structure.

## Disjoint Mapping Network for Cross-modal Matching of Voices and Faces

- Yandong Wen, Mahmoud Al Ismail, Weiyang Liu, Bhiksha Raj, Rita Singh
- abstract@open-review(Poster): We propose a novel framework, called Disjoint Mapping Network (DIMNet), for cross-modal biometric matching, in particular of voices and faces. Different from the existing methods, DIMNet does not explicitly learn the joint relationship between the modalities. Instead, DIMNet learns a shared representation for different modalities by mapping them individually to their common covariates. These shared representations can then be used to find the correspondences between the modalities. We show empirically that DIMNet is able to achieve better performance than the current state-of-the-art methods, with the additional benefits of being conceptually simpler and less data-intensive.

## Learning Procedural Abstractions and Evaluating Discrete Latent Temporal Structure

- Karan Goel, Emma Brunskill
- abstract@open-review(Poster): Clustering methods and latent variable models are often used as tools for pattern mining and discovery of latent structure in time-series data. In this work, we consider the problem of learning procedural abstractions from possibly high-dimensional observational sequences, such as video demonstrations. Given a dataset of time-series, the goal is to identify the latent sequence of steps common to them and label each time-series with the temporal extent of these procedural steps. We introduce a hierarchical Bayesian model called Prism that models the realization of a common procedure across multiple time-series, and can recover procedural abstractions with supervision. We also bring to light two characteristics ignored by traditional evaluation criteria when evaluating latent temporal labelings (temporal clusterings) -- segment structure, and repeated structure -- and develop new metrics tailored to their evaluation. We demonstrate that our metrics improve interpretability and ease of analysis for evaluation on benchmark time-series datasets. Results on benchmark and video datasets indicate that Prism outperforms standard sequence models as well as state-of-the-art techniques in identifying procedural abstractions.

## Learning to Design RNA

- Frederic Runge, Danny Stoll, Stefan Falkner, Frank Hutter

- abstract@open-review(Poster): Designing RNA molecules has garnered recent interest in medicine, synthetic biology, biotechnology and bioinformatics since many functional RNA molecules were shown to be involved in regulatory processes for transcription, epigenetics and translation. Since an RNA's function depends on its structural properties, the RNA Design problem is to find an RNA sequence which satisfies given structural constraints. Here, we propose a new algorithm for the RNA Design problem, dubbed LEARNA. LEARNA uses deep reinforcement learning to train a policy network to sequentially design an entire RNA sequence given a specified target structure. By meta-learning across 65000 different RNA Design tasks for one hour on 20 CPU cores, our extension Meta-LEARNA constructs an RNA Design policy that can be applied out of the box to solve novel RNA Design tasks. Methodologically, for what we believe to be the first time, we jointly optimize over a rich space of architectures for the policy network, the hyperparameters of the training procedure and the formulation of the decision process. Comprehensive empirical results on two widely-used RNA Design benchmarks, as well as a third one that we introduce, show that our approach achieves new state-of-the-art performance on the former while also being orders of magnitudes faster in reaching the previous state-of-the-art performance. In an ablation study, we analyze the importance of our method's different components.

## Cost-Sensitive Robustness against Adversarial Examples

- Xiao Zhang, David Evans
- abstract@open-review(Poster): Several recent works have developed methods for training classifiers that are certifiably robust against norm-bounded adversarial perturbations. These methods assume that all the adversarial transformations are equally important, which is seldom the case in real-world applications. We advocate for cost-sensitive robustness as the criteria for measuring the classifier's performance for tasks where some adversarial transformation are more important than others. We encode the potential harm of each adversarial transformation in a cost matrix, and propose a general objective function to adapt the robust training method of Wong & Kolter (2018) to optimize for cost-sensitive robustness. Our experiments on simple MNIST and CIFAR10 models with a variety of cost matrices show that the proposed approach can produce models with substantially reduced cost-sensitive robust error, while maintaining classification accuracy.

## Combinatorial Attacks on Binarized Neural Networks

- Elias B Khalil, Amrita Gupta, Bistra Dilkina
- abstract@open-review(Poster): Binarized Neural Networks (BNNs) have recently attracted significant interest due to their computational efficiency. Concurrently, it has been shown that neural networks may be overly sensitive to ``attacks" -- tiny adversarial changes in the input -- which may be detrimental to their use in safety-critical domains. Designing attack algorithms that effectively fool trained models is a key step towards learning robust neural networks. The discrete, non-differentiable nature of BNNs, which distinguishes them from their full-precision counterparts, poses a challenge to gradient-based attacks. In this work, we study the problem of attacking a BNN through the lens of combinatorial and integer optimization. We propose a Mixed Integer Linear Programming (MILP) formulation of the problem. While exact and flexible, the MILP quickly becomes intractable as the network and perturbation space grow. To address this issue, we propose IProp, a decomposition-based algorithm that solves a sequence of much smaller MILP problems. Experimentally, we evaluate both proposed methods against the standard gradient-based attack (PGD) on MNIST and Fashion-MNIST, and show that IProp performs favorably compared to PGD, while scaling beyond the limits of the MILP.

## A Variational Inequality Perspective on Generative Adversarial Networks

- Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, Simon Lacoste-Julien
- abstract@open-review(Poster): Generative adversarial networks (GANs) form a generative modeling approach known for producing appealing samples, but they are notably difficult to train. One common way to tackle this issue has been to propose new formulations of the GAN objective. Yet, surprisingly few studies have looked at optimization methods designed for this adversarial training. In this work, we cast GAN optimization problems in the general variational inequality framework. Tapping into the mathematical programming literature, we counter some common misconceptions about the difficulties of saddle point optimization and propose to extend methods designed for variational inequalities to the training of GANs. We apply averaging, extrapolation and a computationally cheaper variant that we call extrapolation from the past to the stochastic gradient method (SGD) and Adam.

## Multiple-Attribute Text Rewriting

- Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, Y-Lan Boureau
- abstract@open-review(Poster): The dominant approach to unsupervised "style transfer" in text is based on the idea of learning a latent representation, which is independent of the attributes specifying its "style". In this paper, we show that this condition is not necessary and is not always met in practice, even with domain adversarial training that explicitly aims at learning such disentangled representations. We thus propose a new model that controls several factors of variation in textual data where this condition on disentanglement is replaced with a simpler mechanism based on back-translation. Our method allows control over multiple attributes, like gender, sentiment, product type, etc., and a more fine-grained control on the trade-off between content preservation and change of style with a pooling operator in the latent space. Our experiments demonstrate that the fully entangled model produces better generations, even when tested on new and more challenging benchmarks comprising reviews with multiple sentences and multiple attributes.

## Multi-Agent Dual Learning

- Yiren Wang, Yingce Xia, Tianyu He, Fei Tian, Tao Qin, ChengXiang Zhai, Tie-Yan Liu
- abstract@open-review(Poster): Dual learning has attracted much attention in machine learning, computer vision and natural language processing communities. The core idea of dual learning is to leverage the duality between the primal task (mapping from domain X to domain Y) and dual task (mapping from domain Y to X) to boost the performances of both tasks. Existing dual learning framework forms a system with two agents (one primal model and one dual model) to utilize such duality. In this paper, we extend this framework by introducing multiple primal and dual models, and propose the multi-agent dual learning framework. Experiments on neural machine translation and image translation tasks demonstrate the effectiveness of the new framework. In particular, we set a new record on IWSLT 2014 German-to-English translation with a 35.44 BLEU score, achieve a 31.03 BLEU score on WMT 2014 English-to-German translation with over 2.6 BLEU improvement over the strong Transformer baseline, and set a new record of 49.61 BLEU score on the recent WMT 2018 English-to-German translation.

## Learning sparse relational transition models

- Victoria Xia, Zi Wang, Kelsey Allen, Tom Silver, Leslie Pack Kaelbling
- abstract@open-review(Poster): We present a representation for describing transition models in complex uncertain domains using relational rules. For any action, a rule selects a set of relevant objects and computes a distribution over properties of just those objects in the resulting state given their properties in the previous state. An iterative greedy algorithm is used to construct a set of deictic references that determine which objects are relevant in any given state. Feed-forward neural networks are used to learn the transition distribution on the relevant objects' properties. This strategy is demonstrated to be both more versatile and more sample efficient than learning a monolithic transition model in a simulated domain in which a robot pushes stacks of objects on a cluttered table.

## SPIGAN: Privileged Adversarial Learning from Simulation

- Kuan-Hui Lee, German Ros, Jie Li, Adrien Gaidon
- abstract@open-review(Poster): Deep Learning for Computer Vision depends mainly on the source of supervision. Photo-realistic simulators can generate large-scale automatically labeled synthetic data, but introduce a domain gap negatively impacting performance. We propose a new unsupervised domain adaptation algorithm, called SPIGAN, relying on Simulator Privileged Information (PI) and Generative Adversarial Networks (GAN). We use internal data from the simulator as PI during the training of a target task network. We experimentally evaluate our approach on semantic segmentation. We train the networks on real-world Cityscapes and Vistas datasets, using only unlabeled real-world images and synthetic labeled data with z-buffer (depth) PI from the SYNTHIA dataset. Our method improves over no adaptation and state-of-the-art unsupervised domain adaptation techniques.

## Universal Stagewise Learning for Non-Convex Problems with Convergence on Averaged Solutions

- Zaiyi Chen, Zhuoning Yuan, Jinfeng Yi, Bowen Zhou, Enhong Chen, Tianbao Yang
- abstract@open-review(Poster): Although stochastic gradient descent (SGD) method and its variants (e.g., stochastic momentum methods, AdaGrad) are algorithms of choice for solving non-convex problems (especially deep learning), big gaps still remain between the theory and the practice with many questions unresolved. For example, there is still a lack of theories of convergence for SGD and its variants that use stagewise step size and return an averaged solution in practice. In addition, theoretical insights of why adaptive step size of AdaGrad could improve non-adaptive step size of SGD is still missing for non-convex optimization. This paper aims to address these questions and fill the gap between theory and practice. We propose a universal stagewise optimization framework for a broad family of non-smooth non-convex problems with the following key features: (i) at each stage any suitable stochastic convex optimization algorithms (e.g., SGD or AdaGrad) that return an averaged solution can be employed for minimizing a regularized convex problem; (ii) the step size is decreased in a stagewise manner; (iii) an averaged solution is returned as the final solution. % that is selected from all stagewise averaged solutions with sampling probabilities increasing as the stage number. Our theoretical results of stagewise {\ada} exhibit its adaptive convergence, therefore shed insights on its faster convergence than stagewise SGD for problems with slowly growing cumulative stochastic gradients. To the best of our knowledge, these new results are the first of their kind for addressing the unresolved issues of existing theories mentioned earlier. Besides theoretical contributions, our empirical studies show that our stagewise variants of SGD, AdaGrad improve the generalization performance of existing variants/implementations of SGD and AdaGrad.

## Reasoning About Physical Interactions with Object-Oriented Prediction and Planning

- Michael Janner, Sergey Levine, William T. Freeman, Joshua B. Tenenbaum, Chelsea Finn, Jiajun Wu
- abstract@open-review(Poster): Object-based factorizations provide a useful level of abstraction for interacting with the world. Building explicit object representations, however, often requires supervisory signals that are difficult to obtain in practice. We present a paradigm for learning object-centric representations for physical scene understanding without direct supervision of object properties. Our model, Object-Oriented Prediction and Planning (O2P2), jointly learns a perception function to map from image observations to object representations, a pairwise physics interaction function to predict the time evolution of a collection of objects, and a rendering function to map objects back to pixels. For evaluation, we consider not only the accuracy of the physical predictions of the model, but also its utility for downstream tasks that require an actionable representation of intuitive physics. After training our model on an image prediction task, we can use its learned representations to build block towers more complicated than those observed during training.

## Posterior Attention Models for Sequence to Sequence Learning

- Shiv Shankar, Sunita Sarawagi
- abstract@open-review(Poster): Modern neural architectures critically rely on attention for mapping structured inputs to sequences. In this paper we show that prevalent attention architectures do not adequately model the dependence among the attention and output tokens across a predicted sequence. We present an alternative architecture called Posterior Attention Models that after a principled factorization of the full joint distribution of the attention and output variables, proposes two major changes. First, the position where attention is marginalized is changed from the input to the output. Second, the attention propagated to the next decoding stage is a posterior attention distribution conditioned on the output. Empirically on five translation and two morphological inflection tasks the proposed posterior attention models yield better BLEU score and alignment accuracy than existing attention models.

## NOODL: Provable Online Dictionary Learning and Sparse Coding

- Sirisha Rambhatla, Xingguo Li, Jarvis Haupt
- abstract@open-review(Poster): We consider the dictionary learning problem, where the aim is to model the given data as a linear combination of a few columns of a matrix known as a dictionary, where the sparse weights forming the linear combination are known as coefficients. Since the dictionary and coefficients, parameterizing the linear model are unknown, the corresponding optimization is inherently non-convex. This was a major challenge until recently, when provable algorithms for dictionary learning were proposed. Yet, these provide guarantees only on the recovery of the dictionary, without explicit recovery guarantees on the coefficients. Moreover, any estimation error in the dictionary adversely impacts the ability to successfully localize and estimate the coefficients. This potentially limits the utility of existing provable dictionary learning methods in applications where coefficient recovery is of interest. To this end, we develop NOODL: a simple Neurally plausible alternating Optimization-based Online Dictionary Learning algorithm, which recovers both the dictionary and coefficients exactly at a geometric rate, when initialized appropriately. Our algorithm, NOODL, is also scalable and amenable for large scale distributed implementations in neural architectures, by which we mean that it only involves simple linear and non-linear operations. Finally, we corroborate these theoretical results via experimental evaluation of the proposed algorithm with the current state-of-the-art techniques.

## RelGAN: Relational Generative Adversarial Networks for Text Generation

- Weili Nie, Nina Narodytska, Ankit Patel
- abstract@open-review(Poster): Generative adversarial networks (GANs) have achieved great success at generating realistic images. However, the text generation still remains a challenging task for modern GAN architectures. In this work, we propose RelGAN, a new GAN architecture for text generation, consisting of three main components: a relational memory based generator for the long-distance dependency modeling, the Gumbel-Softmax relaxation for training GANs on discrete data, and multiple embedded representations in the discriminator to provide a more informative signal for the generator updates. Our experiments show that RelGAN outperforms current state-of-the-art models in terms of sample quality and diversity, and we also reveal via ablation studies that each component of RelGAN contributes critically to its performance improvements. Moreover, a key advantage of our method, that distinguishes it from other GANs, is the ability to control the trade-off between sample quality and diversity via the use of a single adjustable parameter. Finally, RelGAN is the first architecture that makes GANs with Gumbel-Softmax relaxation succeed in generating realistic text.

## Do Deep Generative Models Know What They Don't Know?

- Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, Balaji Lakshminarayanan
- abstract@open-review(Poster): A neural network deployed in the wild may be asked to make predictions for inputs that were drawn from a different distribution than that of the training data. A plethora of work has demonstrated that it is easy to find or synthesize inputs for which a neural network is highly confident yet wrong. Generative models are widely viewed to be robust to such mistaken confidence as modeling the density of the input features can be used to detect novel, out-of-distribution inputs. In this paper we challenge this assumption. We find that the density learned by flow-based models, VAEs, and PixelCNNs cannot distinguish images of common objects such as dogs, trucks, and horses (i.e. CIFAR-10) from those of house numbers (i.e. SVHN), assigning a higher likelihood to the latter when the model is trained on the former. Moreover, we find evidence of this phenomenon when pairing

several popular image data sets: FashionMNIST vs MNIST, CelebA vs SVHN, ImageNet vs CIFAR-10 / CIFAR-100 / SVHN. To investigate this curious behavior, we focus analysis on flow-based generative models in particular since they are trained and evaluated via the exact marginal likelihood. We find such behavior persists even when we restrict the flows to constant-volume transformations. These transformations admit some theoretical analysis, and we show that the difference in likelihoods can be explained by the location and variances of the data and the model curvature. Our results caution against using the density estimates from deep generative models to identify inputs similar to the training distribution until their behavior for out-of-distribution inputs is better understood.

## K for the Price of 1: Parameter-efficient Multi-task and Transfer Learning

- Pramod Kaushik Mudrakarta, Mark Sandler, Andrey Zhmoginov, Andrew Howard
- abstract@open-review(Poster): We introduce a novel method that enables parameter-efficient transfer and multi-task learning with deep neural networks. The basic approach is to learn a model patch - a small set of parameters - that will specialize to each task, instead of fine-tuning the last layer or the entire network. For instance, we show that learning a set of scales and biases is sufficient to convert a pretrained network to perform well on qualitatively different problems (e.g. converting a Single Shot MultiBox Detection (SSD) model into a 1000-class image classification model while reusing 98% of parameters of the SSD feature extractor). Similarly, we show that re-learning existing low-parameter layers (such as depth-wise convolutions) while keeping the rest of the network frozen also improves transfer-learning accuracy significantly. Our approach allows both simultaneous (multi-task) as well as sequential transfer learning. In several multi-task learning problems, despite using much fewer parameters than traditional logits-only fine-tuning, we match single-task performance.

## MisGAN: Learning from Incomplete Data with Generative Adversarial Networks

- Steven Cheng-Xian Li, Bo Jiang, Benjamin Marlin
- abstract@open-review(Poster): Generative adversarial networks (GANs) have been shown to provide an effective way to model complex distributions and have obtained impressive results on various challenging tasks. However, typical GANs require fully-observed data during training. In this paper, we present a GAN-based framework for learning from complex, high-dimensional incomplete data. The proposed framework learns a complete data generator along with a mask generator that models the missing data distribution. We further demonstrate how to impute missing data by equipping our framework with an adversarially trained imputer. We evaluate the proposed framework using a series of experiments with several types of missing data processes under the missing completely at random assumption.

## Learnable Embedding Space for Efficient Neural Architecture Compression

- Shengcao Cao, Xiaofang Wang, Kris M. Kitani
- abstract@open-review(Poster): We propose a method to incrementally learn an embedding space over the domain of network architectures, to enable the careful selection of architectures for evaluation during compressed architecture search. Given a teacher network, we search for a compressed network architecture by using Bayesian Optimization (BO) with a kernel function defined over our proposed embedding space to select architectures for evaluation. We demonstrate that our search algorithm can significantly outperform various baseline methods, such as random search and reinforcement learning (Ashok et al., 2018). The compressed architectures found by our method are also better than the state-of-the-art manually-designed compact architecture ShuffleNet (Zhang et al., 2018). We also demonstrate that the learned embedding space can be transferred to new settings for architecture search, such as a larger teacher network or a teacher network in a different architecture family, without any training.

## Guiding Policies with Language via Meta-Learning

- John D. Co-Reyes, Abhishek Gupta, Suvansh Sanjeev, Nick Altieri, Jacob Andreas, John DeNero, Pieter Abbeel, Sergey Levine
- abstract@open-review(Poster): Behavioral skills or policies for autonomous agents are conventionally learned from reward functions, via reinforcement learning, or from demonstrations, via imitation learning. However, both modes of task specification have their disadvantages: reward functions require manual engineering, while demonstrations require a human expert to be able to actually perform the task in order to generate the demonstration. Instruction following from natural language instructions provides an appealing alternative: in the same way that we can specify goals to other humans simply by speaking or writing, we would like to be able to specify tasks for our machines. However, a single instruction may be insufficient to fully communicate our intent or, even if it is, may be insufficient for an autonomous agent to actually understand how to perform the desired task. In this work, we propose an interactive formulation of the task specification problem, where iterative language corrections are provided to an autonomous agent, guiding it in acquiring the desired skill. Our proposed language-guided policy learning algorithm can integrate an instruction and a sequence of corrections to acquire new skills very quickly. In our experiments, we show that this method can enable a policy to follow instructions and corrections for simulated navigation and manipulation tasks, substantially outperforming direct, non-interactive instruction following.

## Active Learning with Partial Feedback

- Peiyun Hu, Zachary C. Lipton, Anima Anandkumar, Deva Ramanan
- abstract@open-review(Poster): While many active learning papers assume that the learner can simply ask for a label and receive it, real annotation often presents a mismatch between the form of a label (say, one among many classes), and the form of an annotation (typically yes/no binary feedback). To annotate examples corpora for multiclass classification, we might need to ask multiple yes/no questions, exploiting a label hierarchy if one is available. To address this more realistic setting, we propose active learning with partial feedback (ALPF), where the learner must actively choose both which example to label and which binary question to ask. At each step, the learner selects an example, asking if it belongs to a chosen (possibly composite) class. Each answer eliminates some classes, leaving the learner with a partial label. The learner may then either ask more questions about the same example (until an exact label is uncovered) or move on immediately, leaving the first example partially labeled. Active learning with partial labels requires (i) a sampling strategy to choose (example, class) pairs, and (ii) learning from partial labels between rounds. Experiments on Tiny ImageNet demonstrate that our most effective method improves 26% (relative) in top-1 classification accuracy compared to i.i.d. baselines and standard active learners given 30% of the annotation budget that would be required (naively) to annotate the dataset. Moreover, ALPF-learners fully annotate TinyImageNet at 42% lower cost. Surprisingly, we observe that accounting for per-example annotation costs can alter the conventional wisdom that active learners should solicit labels for hard examples.

## On the Sensitivity of Adversarial Robustness to Input Data Distributions

- Gavin Weiguang Ding, Kry Yik Chau Lui, Xiaomeng Jin, Luyu Wang, Ruitong Huang
- abstract@open-review(Poster): Neural networks are vulnerable to small adversarial perturbations. Existing literature largely focused on understanding and mitigating the vulnerability of learned models. In this paper, we demonstrate an intriguing phenomenon about the most popular robust training method in the literature, adversarial training: Adversarial robustness, unlike clean accuracy, is sensitive to the input data distribution. Even a semantics-preserving transformations on the input data distribution can cause a significantly different robustness for the adversarial trained model that is both trained and evaluated on the new distribution. Our discovery of such sensitivity on data distribution is based on a study which disentangles the behaviors of clean accuracy and robust accuracy of the Bayes classifier. Empirical investigations further confirm our finding. We construct semantically-identical variants for MNIST and CIFAR10 respectively, and show that standardly trained models achieve comparable clean accuracies on them, but adversarially trained models achieve significantly different robustness accuracies. This counter-intuitive phenomenon indicates that input data distribution alone can affect the

adversarial robustness of trained neural networks, not necessarily the tasks themselves. Lastly, we discuss the practical implications on evaluating adversarial robustness, and make initial attempts to understand this complex phenomenon.

## Efficient Multi-Objective Neural Architecture Search via Lamarckian Evolution

- Thomas Elsken, Jan Hendrik Metzen, Frank Hutter
- abstract@open-review(Poster): Architecture search aims at automatically finding neural architectures that are competitive with architectures designed by human experts. While recent approaches have achieved state-of-the-art predictive performance for image recognition, they are problematic under resource constraints for two reasons: (1) the neural architectures found are solely optimized for high predictive performance, without penalizing excessive resource consumption; (2)most architecture search methods require vast computational resources. We address the first shortcoming by proposing LEMONADE, an evolutionary algorithm for multi-objective architecture search that allows approximating the Pareto-front of architectures under multiple objectives, such as predictive performance and number of parameters, in a single run of the method. We address the second shortcoming by proposing a Lamarckian inheritance mechanism for LEMONADE which generates children networks that are warmstarted with the predictive performance of their trained parents. This is accomplished by using (approximate) network morphism operators for generating children. The combination of these two contributions allows finding models that are on par or even outperform different-sized NASNets, MobileNets, MobileNets V2 and Wide Residual Networks on CIFAR-10 and ImageNet64x64 within only one week on eight GPUs, which is about 20-40x less compute power than previous architecture search methods that yield state-of-the-art performance.

## DISTRIBUTIONAL CONCAVITY REGULARIZATION FOR GANS

- Shoichiro Yamaguchi, Masanori Koyama
- abstract@open-review(Poster): We propose Distributional Concavity (DC) regularization for Generative Adversarial Networks (GANs), a functional gradient-based method that promotes the entropy of the generator distribution and works against mode collapse. Our DC regularization is an easy-to-implement method that can be used in combination with the current state of the art methods like Spectral Normalization and Wasserstein GAN with gradient penalty to further improve the performance. We will not only show that our DC regularization can achieve highly competitive results on ILSVRC2012 and CIFAR datasets in terms of Inception score and Fr\'echet inception distance, but also provide a mathematical guarantee that our method can always increase the entropy of the generator distribution. We will also show an intimate theoretical connection between our method and the theory of optimal transport.

## Characterizing Audio Adversarial Examples Using Temporal Dependency

- Zhuolin Yang, Bo Li, Pin-Yu Chen, Dawn Song
- abstract@open-review(Poster): Recent studies have highlighted adversarial examples as a ubiquitous threat to different neural network models and many downstream applications. Nonetheless, as unique data properties have inspired distinct and powerful learning principles, this paper aims to explore their potentials towards mitigating adversarial inputs. In particular, our results reveal the importance of using the temporal dependency in audio data to gain discriminate power against adversarial examples. Tested on the automatic speech recognition (ASR) tasks and three recent audio adversarial attacks, we find that (i) input transformation developed from image adversarial defense provides limited robustness improvement and is subtle to advanced attacks; (ii) temporal dependency can be exploited to gain discriminative power against audio adversarial examples and is resistant to adaptive attacks considered in our experiments. Our results not only show promising means of improving the robustness of ASR systems, but also offer novel insights in exploiting domain-specific data properties to mitigate negative effects of adversarial examples.

## GANSynth: Adversarial Neural Audio Synthesis

- Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, Adam Roberts
- abstract@open-review(Poster): Efficient audio synthesis is an inherently difficult machine learning task, as human perception is sensitive to both global structure and fine-scale waveform coherence. Autoregressive models, such as WaveNet, model local structure at the expense of global latent structure and slow iterative sampling, while Generative Adversarial Networks (GANs), have global latent conditioning and efficient parallel sampling, but struggle to generate locally-coherent audio waveforms. Herein, we demonstrate that GANs can in fact generate high-fidelity and locally-coherent audio by modeling log magnitudes and instantaneous frequencies with sufficient frequency resolution in the spectral domain. Through extensive empirical investigations on the NSynth dataset, we demonstrate that GANs are able to outperform strong WaveNet baselines on automated and human evaluation metrics, and efficiently generate audio several orders of magnitude faster than their autoregressive counterparts.

## Toward Understanding the Impact of Staleness in Distributed Machine Learning

- Wei Dai, Yi Zhou, Nanqing Dong, Hao Zhang, Eric Xing
- abstract@open-review(Poster): Most distributed machine learning (ML) systems store a copy of the model parameters locally on each machine to minimize network communication. In practice, in order to reduce synchronization waiting time, these copies of the model are not necessarily updated in lock-step, and can become stale. Despite much development in large-scale ML, the effect of staleness on the learning efficiency is inconclusive, mainly because it is challenging to control or monitor the staleness in complex distributed environments. In this work, we study the convergence behaviors of a wide array of ML models and algorithms under delayed updates. Our extensive experiments reveal the rich diversity of the effects of staleness on the convergence of ML algorithms and offer insights into seemingly contradictory reports in the literature. The empirical findings also inspire a new convergence analysis of SGD in non-convex optimization under staleness, matching the best-known convergence rate of O(1/\sqrt{T}).

## Beyond Greedy Ranking: Slate Optimization via List-CVAE

- Ray Jiang, Sven Gowal, Yuqiu Qian, Timothy Mann, Danilo J. Rezende
- abstract@open-review(Poster): The conventional approach to solving the recommendation problem greedily ranks individual document candidates by prediction scores. However, this method fails to optimize the slate as a whole, and hence, often struggles to capture biases caused by the page layout and document interdepedencies. The slate recommendation problem aims to directly find the optimally ordered subset of documents (i.e. slates) that best serve users' interests. Solving this problem is hard due to the combinatorial explosion of document candidates and their display positions on the page. Therefore we propose a paradigm shift from the traditional viewpoint of solving a ranking problem to a direct slate generation framework. In this paper, we introduce List Conditional Variational Auto-Encoders (ListCVAE), which learn the joint distribution of documents on the slate conditioned on user responses, and directly generate full slates. Experiments on simulated and real-world data show that List-CVAE outperforms greedy ranking methods consistently on various scales of documents corpora.

## G-SGD: Optimizing ReLU Neural Networks in its Positively Scale-Invariant Space

- Qi Meng, Shuxin Zheng, Huishuai Zhang, Wei Chen, Qiwei Ye, Zhi-Ming Ma, Nenghai Yu, Tie-Yan Liu
- abstract@open-review(Poster): It is well known that neural networks with rectified linear units (ReLU) activation functions are positively scale-invariant. Conventional algorithms like stochastic gradient descent optimize the neural networks in the vector space of weights, which is, however, not positively scale-invariant. This mismatch may lead to problems during the optimization process. Then, a natural question is: \emph{can we construct a new vector

space that is positively scale-invariant and sufficient to represent ReLU neural networks so as to better facilitate the optimization process }? In this paper, we provide our positive answer to this question. First, we conduct a formal study on the positive scaling operators which forms a transformation group, denoted as $\mathcal{G}$. We prove that the value of a path (i.e. the product of the weights along the path) in the neural network is invariant to positive scaling and the value vector of all the paths is sufficient to represent the neural networks under mild conditions. Second, we show that one can identify some basis paths out of all the paths and prove that the linear span of their value vectors (denoted as $\mathcal{G}$-space) is an invariant space with lower dimension under the positive scaling group. Finally, we design stochastic gradient descent algorithm in $\mathcal{G}$-space (abbreviated as $\mathcal{G}$-SGD) to optimize the value vector of the basis paths of neural networks with little extra cost by leveraging back-propagation. Our experiments show that $\mathcal{G}$-SGD significantly outperforms the conventional SGD algorithm in optimizing ReLU networks on benchmark datasets.

## Spherical CNNs on Unstructured Grids

- Chiyu Max Jiang, Jingwei Huang, Karthik Kashinath, Prabhat, Philip Marcus, Matthias Niessner
- abstract@open-review(Poster): We present an efficient convolution kernel for Convolutional Neural Networks (CNNs) on unstructured grids using parameterized differential operators while focusing on spherical signals such as panorama images or planetary signals. To this end, we replace conventional convolution kernels with linear combinations of differential operators that are weighted by learnable parameters. Differential operators can be efficiently estimated on unstructured grids using one-ring neighbors, and learnable parameters can be optimized through standard back-propagation. As a result, we obtain extremely efficient neural networks that match or outperform state-of-the-art network architectures in terms of performance but with a significantly lower number of network parameters. We evaluate our algorithm in an extensive series of experiments on a variety of computer vision and climate science tasks, including shape classification, climate pattern segmentation, and omnidirectional image semantic segmentation. Overall, we present (1) a novel CNN approach on unstructured grids using parameterized differential operators for spherical signals, and (2) we show that our unique kernel parameterization allows our model to achieve the same or higher accuracy with significantly fewer network parameters.

## Boosting Robustness Certification of Neural Networks

- Gagandeep Singh, Timon Gehr, Markus Püschel, Martin Vechev
- abstract@open-review(Poster): We present a novel approach for the certification of neural networks against adversarial perturbations which combines scalable overapproximation methods with precise (mixed integer) linear programming. This results in significantly better precision than state-of-the-art verifiers on challenging feedforward and convolutional neural networks with piecewise linear activation functions.

## Two-Timescale Networks for Nonlinear Value Function Approximation

- Wesley Chung, Somjit Nath, Ajin Joseph, Martha White
- abstract@open-review(Poster): A key component for many reinforcement learning agents is to learn a value function, either for policy evaluation or control. Many of the algorithms for learning values, however, are designed for linear function approximation---with a fixed basis or fixed representation. Though there have been a few sound extensions to nonlinear function approximation, such as nonlinear gradient temporal difference learning, these methods have largely not been adopted, eschewed in favour of simpler but not sound methods like temporal difference learning and Q-learning. In this work, we provide a two-timescale network (TTN) architecture that enables linear methods to be used to learn values, with a nonlinear representation learned at a slower timescale. The approach facilitates the use of algorithms developed for the linear setting, such as data-efficient least-squares methods, eligibility traces and the myriad of recently developed linear policy evaluation algorithms, to provide nonlinear value estimates. We prove convergence for TTNs, with particular care given to ensure convergence of the fast linear component under potentially dependent features provided by the learned representation. We empirically demonstrate the benefits of TTNs, compared to other nonlinear value function approximation algorithms, both for policy evaluation and control.

## Differentiable Perturb-and-Parse: Semi-Supervised Parsing with a Structured Variational Autoencoder

- Caio Corro, Ivan Titov
- abstract@open-review(Poster): Human annotation for syntactic parsing is expensive, and large resources are available only for a fraction of languages. A question we ask is whether one can leverage abundant unlabeled texts to improve syntactic parsers, beyond just using the texts to obtain more generalisable lexical features (i.e. beyond word embeddings). To this end, we propose a novel latent-variable generative model for semi-supervised syntactic dependency parsing. As exact inference is intractable, we introduce a differentiable relaxation to obtain approximate samples and compute gradients with respect to the parser parameters. Our method (Differentiable Perturb-and-Parse) relies on differentiable dynamic programming over stochastically perturbed edge scores. We demonstrate effectiveness of our approach with experiments on English, French and Swedish.

## Algorithmic Framework for Model-based Deep Reinforcement Learning with Theoretical Guarantees

- Yuping Luo, Huazhe Xu, Yuanzhi Li, Yuandong Tian, Trevor Darrell, Tengyu Ma
- abstract@open-review(Poster): Model-based reinforcement learning (RL) is considered to be a promising approach to reduce the sample complexity that hinders model-free RL. However, the theoretical understanding of such methods has been rather limited. This paper introduces a novel algorithmic framework for designing and analyzing model-based RL algorithms with theoretical guarantees. We design a meta-algorithm with a theoretical guarantee of monotone improvement to a local maximum of the expected reward. The meta-algorithm iteratively builds a lower bound of the expected reward based on the estimated dynamical model and sample trajectories, and then maximizes the lower bound jointly over the policy and the model. The framework extends the optimism-in-face-of-uncertainty principle to non-linear dynamical models in a way that requires no explicit uncertainty quantification. Instantiating our framework with simplification gives a variant of model-based RL algorithms Stochastic Lower Bounds Optimization (SLBO). Experiments demonstrate that SLBO achieves the state-of-the-art performance when only 1M or fewer samples are permitted on a range of continuous control benchmark tasks.

## Tree-Structured Recurrent Switching Linear Dynamical Systems for Multi-Scale Modeling

- Josue Nassar, Scott Linderman, Monica Bugallo, Il Memming Park
- abstract@open-review(Poster): Many real-world systems studied are governed by complex, nonlinear dynamics. By modeling these dynamics, we can gain insight into how these systems work, make predictions about how they will behave, and develop strategies for controlling them. While there are many methods for modeling nonlinear dynamical systems, existing techniques face a trade off between offering interpretable descriptions and making accurate predictions. Here, we develop a class of models that aims to achieve both simultaneously, smoothly interpolating between simple descriptions and more complex, yet also more accurate models. Our probabilistic model achieves this multi-scale property through of a hierarchy of locally linear dynamics that jointly approximate global nonlinear dynamics. We call it the tree-structured recurrent switching linear dynamical system. To fit this model, we present a fully-Bayesian sampling procedure using Polya-Gamma data augmentation to allow for fast and conjugate Gibbs sampling. Through a variety of synthetic and real examples, we show how these models outperform existing methods in both interpretability and predictive capability.

## Diagnosing and Enhancing VAE Models

- Bin Dai, David Wipf
- abstract@open-review(Poster): Although variational autoencoders (VAEs) represent a widely influential deep generative model, many aspects of the underlying energy function remain poorly understood. In particular, it is commonly believed that Gaussian encoder/decoder assumptions reduce the effectiveness of VAEs in generating realistic samples. In this regard, we rigorously analyze the VAE objective, differentiating situations where this belief is and is not actually true. We then leverage the corresponding insights to develop a simple VAE enhancement that requires no additional hyperparameters or sensitive tuning. Quantitatively, this proposal produces crisp samples and stable FID scores that are actually competitive with a variety of GAN models, all while retaining desirable attributes of the original VAE architecture. The code for our model is available at \url{https://github.com/daib13/TwoStageVAE}.

## Efficiently testing local optimality and escaping saddles for ReLU networks

- Chulhee Yun, Suvrit Sra, Ali Jadbabaie
- abstract@open-review(Poster): We provide a theoretical algorithm for checking local optimality and escaping saddles at nondifferentiable points of empirical risks of two-layer ReLU networks. Our algorithm receives any parameter value and returns: local minimum, second-order stationary point, or a strict descent direction. The presence of M data points on the nondifferentiability of the ReLU divides the parameter space into at most 2^M regions, which makes analysis difficult. By exploiting polyhedral geometry, we reduce the total computation down to one convex quadratic program (QP) for each hidden node, O(M) (in)equality tests, and one (or a few) nonconvex QP. For the last QP, we show that our specific problem can be solved efficiently, in spite of nonconvexity. In the benign case, we solve one equality constrained QP, and we prove that projected gradient descent solves it exponentially fast. In the bad case, we have to solve a few more inequality constrained QPs, but we prove that the time complexity is exponential only in the number of inequality constraints. Our experiments show that either benign case or bad case with very few inequality constraints occurs, implying that our algorithm is efficient in most cases.

## Don't let your Discriminator be fooled

- Brady Zhou, Philipp Krähenbühl
- abstract@open-review(Poster): Generative Adversarial Networks are one of the leading tools in generative modeling, image editing and content creation. However, they are hard to train as they require a delicate balancing act between two deep networks fighting a never ending duel. Some of the most promising adversarial models today minimize a Wasserstein objective. It is smoother and more stable to optimize. In this paper, we show that the Wasserstein distance is just one out of a large family of objective functions that yield these properties. By making the discriminator of a GAN robust to adversarial attacks we can turn any GAN objective into a smooth and stable loss. We experimentally show that any GAN objective, including Wasserstein GANs, benefit from adversarial robustness both quantitatively and qualitatively. The training additionally becomes more robust to suboptimal choices of hyperparameters, model architectures, or objective functions.

## Rigorous Agent Evaluation: An Adversarial Approach to Uncover Catastrophic Failures

- Jonathan Uesato, *Ananya Kumar*, Csaba Szepesvari*, Tom Erez, Avraham Ruderman, Keith Anderson, Krishnamurthy (Dj) Dvijotham, Nicolas Heess, Pushmeet Kohli
- abstract@open-review(Poster): This paper addresses the problem of evaluating learning systems in safety critical domains such as autonomous driving, where failures can have catastrophic consequences. We focus on two problems: searching for scenarios when learned agents fail and assessing their probability of failure. The standard method for agent evaluation in reinforcement learning, Vanilla Monte Carlo, can miss failures entirely, leading to the deployment of unsafe agents. We demonstrate this is an issue for current agents, where even matching the compute used for training is sometimes insufficient for evaluation. To address this shortcoming, we draw upon the rare event probability estimation literature and propose an adversarial evaluation approach. Our approach focuses evaluation on adversarially chosen situations, while still providing unbiased estimates of failure probabilities. The key difficulty is in identifying these adversarial situations -- since failures are rare there is little signal to drive optimization. To solve this we propose a continuation approach that learns failure modes in related but less robust agents. Our approach also allows reuse of data already collected for training the agent. We demonstrate the efficacy of adversarial evaluation on two standard domains: humanoid control and simulated driving. Experimental results show that our methods can find catastrophic failures and estimate failures rates of agents multiple orders of magnitude faster than standard evaluation schemes, in minutes to hours rather than days.

## Competitive experience replay

- Hao Liu, Alexander Trott, Richard Socher, Caiming Xiong
- abstract@open-review(Poster): Deep learning has achieved remarkable successes in solving challenging reinforcement learning (RL) problems when dense reward function is provided. However, in sparse reward environment it still often suffers from the need to carefully shape reward function to guide policy optimization. This limits the applicability of RL in the real world since both reinforcement learning and domain-specific knowledge are required. It is therefore of great practical importance to develop algorithms which can learn from a binary signal indicating successful task completion or other unshaped, sparse reward signals. We propose a novel method called competitive experience replay, which efficiently supplements a sparse reward by placing learning in the context of an exploration competition between a pair of agents. Our method complements the recently proposed hindsight experience replay (HER) by inducing an automatic exploratory curriculum. We evaluate our approach on the tasks of reaching various goal locations in an ant maze and manipulating objects with a robotic arm. Each task provides only binary rewards indicating whether or not the goal is achieved. Our method asymmetrically augments these sparse rewards for a pair of agents each learning the same task, creating a competitive game designed to drive exploration. Extensive experiments demonstrate that this method leads to faster converge and improved task performance.

## A Max-Affine Spline Perspective of Recurrent Neural Networks

- Zichao Wang, Randall Balestriero, Richard Baraniuk
- abstract@open-review(Poster): We develop a framework for understanding and improving recurrent neural networks (RNNs) using max-affine spline operators (MASOs). We prove that RNNs using piecewise affine and convex nonlinearities can be written as a simple piecewise affine spline operator. The resulting representation provides several new perspectives for analyzing RNNs, three of which we study in this paper. First, we show that an RNN internally partitions the input space during training and that it builds up the partition through time. Second, we show that the affine slope parameter of an RNN corresponds to an input-specific template, from which we can interpret an RNN as performing a simple template matching (matched filtering) given the input. Third, by carefully examining the MASO RNN affine mapping, we prove that using a random initial hidden state corresponds to an explicit L2 regularization of the affine parameters, which can mollify exploding gradients and improve generalization. Extensive experiments on several datasets of various modalities demonstrate and validate each of the above conclusions. In particular, using a random initial hidden states elevates simple RNNs to near state-of-the-art performers on these datasets.

## Top-Down Neural Model For Formulae

- Karel Chvalovský
- abstract@open-review(Poster): We present a simple neural model that given a formula and a property tries to answer the question whether the formula has the given property, for example whether a propositional formula is always true. The structure of the formula is captured by a feedforward neural network recursively built for the given formula in a top-down manner. The results of this network are then processed by two recurrent neural networks. One of the

interesting aspects of our model is how propositional atoms are treated. For example, the model is insensitive to their names, it only matters whether they are the same or distinct.

## Feature-Wise Bias Amplification

- Klas Leino, Emily Black, Matt Fredrikson, Shayak Sen, Anupam Datta
- abstract@open-review(Poster): We study the phenomenon of bias amplification in classifiers, wherein a machine learning model learns to predict classes with a greater disparity than the underlying ground truth. We demonstrate that bias amplification can arise via inductive bias in gradient descent methods resulting in overestimation of importance of moderately-predictive ``weak'' features if insufficient training data is available. This overestimation gives rise to feature-wise bias amplification -- a previously unreported form of bias that can be traced back to the features of a trained model. Through analysis and experiments, we show that the while some bias cannot be mitigated without sacrificing accuracy, feature-wise bias amplification can be mitigated through targeted feature selection. We present two new feature selection algorithms for mitigating bias amplification in linear models, and show how they can be adapted to convolutional neural networks efficiently. Our experiments on synthetic and real data demonstrate that these algorithms consistently lead to reduced bias without harming accuracy, in some cases eliminating predictive bias altogether while providing modest gains in accuracy.

## AD-VAT: An Asymmetric Dueling mechanism for learning Visual Active Tracking

- Fangwei Zhong, Peng Sun, Wenhan Luo, Tingyun Yan, Yizhou Wang
- abstract@open-review(Poster): Visual Active Tracking (VAT) aims at following a target object by autonomously controlling the motion system of a tracker given visual observations. Previous work has shown that the tracker can be trained in a simulator via reinforcement learning and deployed in real-world scenarios. However, during training, such a method requires manually specifying the moving path of the target object to be tracked, which cannot ensure the tracker’s generalization on the unseen object moving patterns. To learn a robust tracker for VAT, in this paper, we propose a novel adversarial RL method which adopts an Asymmetric Dueling mechanism, referred to as AD-VAT. In AD-VAT, both the tracker and the target are approximated by end-to-end neural networks, and are trained via RL in a dueling/competitive manner: i.e., the tracker intends to lockup the target, while the target tries to escape from the tracker. They are asymmetric in that the target is aware of the tracker, but not vice versa. Specifically, besides its own observation, the target is fed with the tracker’s observation and action, and learns to predict the tracker’s reward as an auxiliary task. We show that such an asymmetric dueling mechanism produces a stronger target, which in turn induces a more robust tracker. To stabilize the training, we also propose a novel partial zero-sum reward for the tracker/target. The experimental results, in both 2D and 3D environments, demonstrate that the proposed method leads to a faster convergence in training and yields more robust tracking behaviors in different testing scenarios. For supplementary videos, see: https://www.youtube.com/playlist?list=PL9rZj4Mea7wOZkdajK1TsprRg8iUf51BS The code is available at https://github.com/zfw1226/active_tracking_rl

## Learning to Learn with Conditional Class Dependencies

- Xiang Jiang, Mohammad Havaei, Farshid Varno, Gabriel Chartrand, Nicolas Chapados, Stan Matwin
- abstract@open-review(Poster): Neural networks can learn to extract statistical properties from data, but they seldom make use of structured information from the label space to help representation learning. Although some label structure can implicitly be obtained when training on huge amounts of data, in a few-shot learning context where little data is available, making explicit use of the label structure can inform the model to reshape the representation space to reflect a global sense of class dependencies. We propose a meta-learning framework, Conditional class-Aware Meta-Learning (CAML), that conditionally transforms feature representations based on a metric space that is trained to capture inter-class dependencies. This enables a conditional modulation of the feature representations of the base-learner to impose regularities informed by the label space. Experiments show that the conditional transformation in CAML leads to more disentangled representations and achieves competitive results on the miniImageNet benchmark.

## GAN Dissection: Visualizing and Understanding Generative Adversarial Networks

- David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, Antonio Torralba
- abstract@open-review(Poster): Generative Adversarial Networks (GANs) have recently achieved impressive results for many real-world applications, and many GAN variants have emerged with improvements in sample quality and training stability. However, visualization and understanding of GANs is largely missing. How does a GAN represent our visual world internally? What causes the artifacts in GAN results? How do architectural choices affect GAN learning? Answering such questions could enable us to develop new insights and better models.

In this work, we present an analytic framework to visualize and understand GANs at the unit-, object-, and scene-level. We first identify a group of interpretable units that are closely related to object concepts with a segmentation-based network dissection method. Then, we quantify the causal effect of interpretable units by measuring the ability of interventions to control objects in the output. Finally, we examine the contextual relationship between these units and their surrounding by inserting the discovered object concepts into new images. We show several practical applications enabled by our framework, from comparing internal representations across different layers, models, and datasets, to improving GANs by locating and removing artifact-causing units, to interactively manipulating objects in the scene. We provide open source interpretation tools to help peer researchers and practitioners better understand their GAN models.

## TimbreTron: A WaveNet(CycleGAN(CQT(Audio))) Pipeline for Musical Timbre Transfer

- Sicong Huang, Qiyang Li, Cem Anil, Xuchan Bao, Sageev Oore, Roger B. Grosse
- abstract@open-review(Poster): In this work, we address the problem of musical timbre transfer, where the goal is to manipulate the timbre of a sound sample from one instrument to match another instrument while preserving other musical content, such as pitch, rhythm, and loudness. In principle, one could apply image-based style transfer techniques to a time-frequency representation of an audio signal, but this depends on having a representation that allows independent manipulation of timbre as well as high-quality waveform generation. We introduce TimbreTron, a method for musical timbre transfer which applies “image” domain style transfer to a time-frequency representation of the audio signal, and then produces a high-quality waveform using a conditional WaveNet synthesizer. We show that the Constant Q Transform (CQT) representation is particularly well-suited to convolutional architectures due to its approximate pitch equivariance. Based on human perceptual evaluations, we confirmed that TimbreTron recognizably transferred the timbre while otherwise preserving the musical content, for both monophonic and polyphonic samples. We made an accompanying demo video here: https://www.cs.toronto.edu/~huang/TimbreTron/index.html which we strongly encourage you to watch before reading the paper.

## A Mean Field Theory of Batch Normalization

- Greg Yang, Jeffrey Pennington, Vinay Rao, Jascha Sohl-Dickstein, Samuel S. Schoenholz
- abstract@open-review(Poster): We develop a mean field theory for batch normalization in fully-connected feedforward neural networks. In so doing, we provide a precise characterization of signal propagation and gradient backpropagation in wide batch-normalized networks at initialization. Our theory shows that gradient signals grow exponentially in depth and that these exploding gradients cannot be eliminated by tuning the initial weight variances or by adjusting the nonlinear activation function. Indeed, batch normalization itself is the cause of gradient explosion. As a result, vanilla batch-normalized networks without skip connections are not trainable at large depths for common initialization schemes, a prediction that we verify with a variety of empirical simulations. While gradient explosion cannot be eliminated, it can be reduced by tuning the network close to the linear regime, which improves the trainability of deep batch-normalized networks without residual connections. Finally, we investigate the learning dynamics of batch-normalized networks and observe that after a single step of optimization the networks achieve a relatively stable equilibrium in which gradients have dramatically

smaller dynamic range. Our theory leverages Laplace, Fourier, and Gegenbauer transforms and we derive new identities that may be of independent interest.

## A Closer Look at Few-shot Classification

- Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, Jia-Bin Huang
- abstract@open-review(Poster): Few-shot classification aims to learn a classifier to recognize unseen classes during training with limited labeled examples. While significant progress has been made, the growing complexity of network designs, meta-learning algorithms, and differences in implementation details make a fair comparison difficult. In this paper, we present 1) a consistent comparative analysis of several representative few-shot classification algorithms, with results showing that deeper backbones significantly reduce the gap across methods including the baseline, 2) a slightly modified baseline method that surprisingly achieves competitive performance when compared with the state-of-the-art on both the mini-ImageNet and the CUB datasets, and 3) a new experimental setting for evaluating the cross-domain generalization ability for few-shot classification algorithms. Our results reveal that reducing intra-class variation is an important factor when the feature backbone is shallow, but not as critical when using deeper backbones. In a realistic, cross-domain evaluation setting, we show that a baseline method with a standard fine-tuning practice compares favorably against other state-of-the-art few-shot learning algorithms.

## STCN: Stochastic Temporal Convolutional Networks

- Emre Aksan, Otmar Hilliges
- abstract@open-review(Poster): Convolutional architectures have recently been shown to be competitive on many sequence modelling tasks when compared to the de-facto standard of recurrent neural networks (RNNs) while providing computational and modelling advantages due to inherent parallelism. However, currently, there remains a performance gap to more expressive stochastic RNN variants, especially those with several layers of dependent random variables. In this work, we propose stochastic temporal convolutional networks (STCNs), a novel architecture that combines the computational advantages of temporal convolutional networks (TCN) with the representational power and robustness of stochastic latent spaces. In particular, we propose a hierarchy of stochastic latent variables that captures temporal dependencies at different time-scales. The architecture is modular and flexible due to the decoupling of the deterministic and stochastic layers. We show that the proposed architecture achieves state of the art log-likelihoods across several tasks. Finally, the model is capable of predicting high-quality synthetic samples over a long-range temporal horizon in modelling of handwritten text.

## RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space

- Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, Jian Tang
- abstract@open-review(Poster): We study the problem of learning representations of entities and relations in knowledge graphs for predicting missing links. The success of such a task heavily relies on the ability of modeling and inferring the patterns of (or between) the relations. In this paper, we present a new approach for knowledge graph embedding called RotatE, which is able to model and infer various relation patterns including: symmetry/antisymmetry, inversion, and composition. Specifically, the RotatE model defines each relation as a rotation from the source entity to the target entity in the complex vector space. In addition, we propose a novel self-adversarial negative sampling technique for efficiently and effectively training the RotatE model. Experimental results on multiple benchmark knowledge graphs show that the proposed RotatE model is not only scalable, but also able to infer and model various relation patterns and significantly outperform existing state-of-the-art models for link prediction.

## Learning to Navigate the Web

- Izzeddin Gur, Ulrich Rueckert, Aleksandra Faust, Dilek Hakkani-Tur
- abstract@open-review(Poster): Learning in environments with large state and action spaces, and sparse rewards, can hinder a Reinforcement Learning (RL) agent’s learning through trial-and-error. For instance, following natural language instructions on the Web (such as booking a flight ticket) leads to RL settings where input vocabulary and number of actionable elements on a page can grow very large. Even though recent approaches improve the success rate on relatively simple environments with the help of human demonstrations to guide the exploration, they still fail in environments where the set of possible instructions can reach millions. We approach the aforementioned problems from a different perspective and propose guided RL approaches that can generate unbounded amount of experience for an agent to learn from. Instead of learning from a complicated instruction with a large vocabulary, we decompose it into multiple sub-instructions and schedule a curriculum in which an agent is tasked with a gradually increasing subset of these relatively easier sub-instructions. In addition, when the expert demonstrations are not available, we propose a novel meta-learning framework that generates new instruction following tasks and trains the agent more effectively. We train DQN, deep reinforcement learning agent, with Q-value function approximated with a novel QWeb neural network architecture on these smaller, synthetic instructions. We evaluate the ability of our agent to generalize to new instructions onWorld of Bits benchmark, on forms with up to 100 elements, supporting 14 million possible instructions. The QWeb agent outperforms the baseline without using any human demonstration achieving 100% success rate on several difficult environments.

## Modeling Uncertainty with Hedged Instance Embeddings

- Seong Joon Oh, Kevin P. Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, Andrew C. Gallagher
- abstract@open-review(Poster): Instance embeddings are an efficient and versatile image representation that facilitates applications like recognition, verification, retrieval, and clustering. Many metric learning methods represent the input as a single point in the embedding space. Often the distance between points is used as a proxy for match confidence. However, this can fail to represent uncertainty which can arise when the input is ambiguous, e.g., due to occlusion or blurriness. This work addresses this issue and explicitly models the uncertainty by “hedging” the location of each input in the embedding space. We introduce the hedged instance embedding (HIB) in which embeddings are modeled as random variables and the model is trained under the variational information bottleneck principle (Alemi et al., 2016; Achille & Soatto, 2018). Empirical results on our new N-digit MNIST dataset show that our method leads to the desired behavior of “hedging its bets” across the embedding space upon encountering ambiguous inputs. This results in improved performance for image matching and classification tasks, more structure in the learned embedding space, and an ability to compute a per-exemplar uncertainty measure which is correlated with downstream performance.

## Automatically Composing Representation Transformations as a Means for Generalization

- Michael Chang, Abhishek Gupta, Sergey Levine, Thomas L. Griffiths
- abstract@open-review(Poster): A generally intelligent learner should generalize to more complex tasks than it has previously encountered, but the two common paradigms in machine learning -- either training a separate learner per task or training a single learner for all tasks -- both have difficulty with such generalization because they do not leverage the compositional structure of the task distribution. This paper introduces the compositional problem graph as a broadly applicable formalism to relate tasks of different complexity in terms of problems with shared subproblems. We propose the compositional generalization problem for measuring how readily old knowledge can be reused and hence built upon. As a first step for tackling compositional generalization, we introduce the compositional recursive learner, a domain-general framework for learning algorithmic procedures for composing representation transformations, producing a learner that reasons about what computation to execute by making analogies to previously seen problems. We show on a symbolic and a high-dimensional domain that our compositional approach can generalize to more complex problems than the learner has previously encountered, whereas baselines that are not explicitly compositional do not.

## Systematic Generalization: What Is Required and Can It Be Learned?

- Dzmitry Bahdanau, *Shikhar Murty*, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, Aaron Courville
- abstract@open-review(Poster): Numerous models for grounded language understanding have been recently proposed, including (i) generic models that can be easily adapted to any given task and (ii) intuitively appealing modular models that require background knowledge to be instantiated. We compare both types of models in how much they lend themselves to a particular form of systematic generalization. Using a synthetic VQA test, we evaluate which models are capable of reasoning about all possible object pairs after training on only a small subset of them. Our findings show that the generalization of modular models is much more systematic and that it is highly sensitive to the module layout, i.e. to how exactly the modules are connected. We furthermore investigate if modular models that generalize well could be made more end-to-end by learning their layout and parametrization. We find that end-to-end methods from prior work often learn inappropriate layouts or parametrizations that do not facilitate systematic generalization. Our results suggest that, in addition to modularity, systematic generalization in language understanding may require explicit regularizers or priors.

## Adaptive Input Representations for Neural Language Modeling

- Alexei Baevski, Michael Auli
- abstract@open-review(Poster): We introduce adaptive input representations for neural language modeling which extend the adaptive softmax of Grave et al. (2017) to input representations of variable capacity. There are several choices on how to factorize the input and output layers, and whether to model words, characters or sub-word units. We perform a systematic comparison of popular choices for a self-attentional architecture. Our experiments show that models equipped with adaptive embeddings are more than twice as fast to train than the popular character input CNN while having a lower number of parameters. On the WikiText-103 benchmark we achieve 18.7 perplexity, an improvement of 10.5 perplexity compared to the previously best published result and on the Billion Word benchmark, we achieve 23.02 perplexity.

## Optimal Control Via Neural Networks: A Convex Approach

- Yize Chen, Yuanyuan Shi, Baosen Zhang
- abstract@open-review(Poster): Control of complex systems involves both system identification and controller design. Deep neural networks have proven to be successful in many identification tasks, however, from model-based control perspective, these networks are difficult to work with because they are typically nonlinear and nonconvex. Therefore many systems are still identified and controlled based on simple linear models despite their poor representation capability.

In this paper we bridge the gap between model accuracy and control tractability faced by neural networks, by explicitly constructing networks that are convex with respect to their inputs. We show that these input convex networks can be trained to obtain accurate models of complex physical systems. In particular, we design input convex recurrent neural networks to capture temporal behavior of dynamical systems. Then optimal controllers can be achieved via solving a convex model predictive control problem. Experiment results demonstrate the good potential of the proposed input convex neural network based approach in a variety of control applications. In particular we show that in the MuJoCo locomotion tasks, we could achieve over 10% higher performance using 5 times less time compared with state-of-the-art model-based reinforcement learning method; and in the building HVAC control example, our method achieved up to 20% energy reduction compared with classic linear models.

## An Empirical Study of Example Forgetting during Deep Neural Network Learning

- Mariya Toneva, *Alessandro Sordoni*, Remi Tachet des Combes*, Adam Trischler, Yoshua Bengio, Geoffrey J. Gordon
- abstract@open-review(Poster): Inspired by the phenomenon of catastrophic forgetting, we investigate the learning dynamics of neural networks as they train on single classification tasks. Our goal is to understand whether a related phenomenon occurs when data does not undergo a clear distributional shift. We define a ``forgetting event'' to have occurred when an individual training example transitions from being classified correctly to incorrectly over the course of learning. Across several benchmark data sets, we find that: (i) certain examples are forgotten with high frequency, and some not at all; (ii) a data set's (un)forgettable examples generalize across neural architectures; and (iii) based on forgetting dynamics, a significant fraction of examples can be omitted from the training data set while still maintaining state-of-the-art generalization performance.

## GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding

- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, Samuel R. Bowman
- abstract@open-review(Poster): For natural language understanding (NLU) technology to be maximally useful, it must be able to process language in a way that is not exclusive to a single task, genre, or dataset. In pursuit of this objective, we introduce the General Language Understanding Evaluation (GLUE) benchmark, a collection of tools for evaluating the performance of models across a diverse set of existing NLU tasks. By including tasks with limited training data, GLUE is designed to favor and encourage models that share general linguistic knowledge across tasks. GLUE also includes a hand-crafted diagnostic test suite that enables detailed linguistic analysis of models. We evaluate baselines based on current methods for transfer and representation learning and find that multi-task training on all tasks performs better than training a separate model per task. However, the low absolute performance of our best model indicates the need for improved general NLU systems.

## Plan Online, Learn Offline: Efficient Learning and Exploration via Model-Based Control

- Kendall Lowrey, Aravind Rajeswaran, Sham Kakade, Emanuel Todorov, Igor Mordatch
- abstract@open-review(Poster): We propose a "plan online and learn offline" framework for the setting where an agent, with an internal model, needs to continually act and learn in the world. Our work builds on the synergistic relationship between local model-based control, global value function learning, and exploration. We study how local trajectory optimization can cope with approximation errors in the value function, and can stabilize and accelerate value function learning. Conversely, we also study how approximate value functions can help reduce the planning horizon and allow for better policies beyond local solutions. Finally, we also demonstrate how trajectory optimization can be used to perform temporally coordinated exploration in conjunction with estimating uncertainty in value function approximation. This exploration is critical for fast and stable learning of the value function. Combining these components enable solutions to complex control tasks, like humanoid locomotion and dexterous in-hand manipulation, in the equivalent of a few minutes of experience in the real world.

## Learning Grid Cells as Vector Representation of Self-Position Coupled with Matrix Representation of Self-Motion

- Ruiqi Gao, Jianwen Xie, Song-Chun Zhu, Ying Nian Wu
- abstract@open-review(Poster): This paper proposes a representational model for grid cells. In this model, the 2D self-position of the agent is represented by a high-dimensional vector, and the 2D self-motion or displacement of the agent is represented by a matrix that transforms the vector. Each component of the vector is a unit or a cell. The model consists of the following three sub-models. (1) Vector-matrix multiplication. The movement from the current position to the next position is modeled by matrix-vector multi- plication, i.e., the vector of the next position is obtained by multiplying the matrix of the motion to the vector of the current position. (2) Magnified local isometry. The angle between two nearby vectors equals the Euclidean distance between the two corresponding positions multiplied by a magnifying factor. (3) Global adjacency kernel. The inner product between two vectors measures the adjacency between the two corresponding positions, which is defined by a kernel function of the Euclidean distance between the two positions. Our

representational model has explicit algebra and geometry. It can learn hexagon patterns of grid cells, and it is capable of error correction, path integral and path planning.

## Preventing Posterior Collapse with delta-VAEs

- Ali Razavi, Aaron van den Oord, Ben Poole, Oriol Vinyals
- abstract@open-review(Poster): Due to the phenomenon of "posterior collapse," current latent variable generative models pose a challenging design choice that either weakens the capacity of the decoder or requires altering the training objective. We develop an alternative that utilizes the most powerful generative models as decoders, optimize the variational lower bound, and ensures that the latent variables preserve and encode useful information. Our proposed δ-VAEs achieve this by constraining the variational family for the posterior to have a minimum distance to the prior. For sequential latent variable models, our approach resembles the classic representation learning approach of slow feature analysis. We demonstrate our method's efficacy at modeling text on LM1B and modeling images: learning representations, improving sample quality, and achieving state of the art log-likelihood on CIFAR-10 and ImageNet 32 × 32.

## Deep Online Learning Via Meta-Learning: Continual Adaptation for Model-Based RL

- Anusha Nagabandi, Chelsea Finn, Sergey Levine
- abstract@open-review(Poster): Humans and animals can learn complex predictive models that allow them to accurately and reliably reason about real-world phenomena, and they can adapt such models extremely quickly in the face of unexpected changes. Deep neural network models allow us to represent very complex functions, but lack this capacity for rapid online adaptation. The goal in this paper is to develop a method for continual online learning from an incoming stream of data, using deep neural network models. We formulate an online learning procedure that uses stochastic gradient descent to update model parameters, and an expectation maximization algorithm with a Chinese restaurant process prior to develop and maintain a mixture of models to handle non-stationary task distributions. This allows for all models to be adapted as necessary, with new models instantiated for task changes and old models recalled when previously seen tasks are encountered again. Furthermore, we observe that meta-learning can be used to meta-train a model such that this direct online adaptation with SGD is effective, which is otherwise not the case for large function approximators. We apply our method to model-based reinforcement learning, where adapting the predictive model is critical for control; we demonstrate that our online learning via meta-learning algorithm outperforms alternative prior methods, and enables effective continuous adaptation in non-stationary task distributions such as varying terrains, motor failures, and unexpected disturbances.

## Value Propagation Networks

- Nantas Nardelli, Gabriel Synnaeve, Zeming Lin, Pushmeet Kohli, Philip H. S. Torr, Nicolas Usunier
- abstract@open-review(Poster): We present Value Propagation (VProp), a set of parameter-efficient differentiable planning modules built on Value Iteration which can successfully be trained using reinforcement learning to solve unseen tasks, has the capability to generalize to larger map sizes, and can learn to navigate in dynamic environments. We show that the modules enable learning to plan when the environment also includes stochastic elements, providing a cost-efficient learning system to build low-level size-invariant planners for a variety of interactive navigation problems. We evaluate on static and dynamic configurations of MazeBase grid-worlds, with randomly generated environments of several different sizes, and on a StarCraft navigation scenario, with more complex dynamics, and pixels as input.

## Efficient Augmentation via Data Subsampling

- Michael Kuchnik, Virginia Smith
- abstract@open-review(Poster): Data augmentation is commonly used to encode invariances in learning methods. However, this process is often performed in an inefficient manner, as artificial examples are created by applying a number of transformations to all points in the training set. The resulting explosion of the dataset size can be an issue in terms of storage and training costs, as well as in selecting and tuning the optimal set of transformations to apply. In this work, we demonstrate that it is possible to significantly reduce the number of data points included in data augmentation while realizing the same accuracy and invariance benefits of augmenting the entire dataset. We propose a novel set of subsampling policies, based on model influence and loss, that can achieve a 90% reduction in augmentation set size while maintaining the accuracy gains of standard data augmentation.

## ALISTA: Analytic Weights Are As Good As Learned Weights in LISTA

- Jialin Liu, Xiaohan Chen, Zhangyang Wang, Wotao Yin
- abstract@open-review(Poster): Deep neural networks based on unfolding an iterative algorithm, for example, LISTA (learned iterative shrinkage thresholding algorithm), have been an empirical success for sparse signal recovery. The weights of these neural networks are currently determined by data-driven "black-box" training. In this work, we propose Analytic LISTA (ALISTA), where the weight matrix in LISTA is computed as the solution to a data-free optimization problem, leaving only the stepsize and threshold parameters to data-driven learning. This significantly simplifies the training. Specifically, the data-free optimization problem is based on coherence minimization. We show our ALISTA retains the optimal linear convergence proved in (Chen et al., 2018) and has a performance comparable to LISTA. Furthermore, we extend ALISTA to convolutional linear operators, again determined in a data-free manner. We also propose a feed-forward framework that combines the data-free optimization and ALISTA networks from end to end, one that can be jointly trained to gain robustness to small perturbations in the encoding model.

## Recall Traces: Backtracking Models for Efficient Reinforcement Learning

- Anirudh Goyal, Philemon Brakel, William Fedus, Soumye Singhal, Timothy Lillicrap, Sergey Levine, Hugo Larochelle, Yoshua Bengio
- abstract@open-review(Poster): In many environments only a tiny subset of all states yield high reward. In these cases, few of the interactions with the environment provide a relevant learning signal. Hence, we may want to preferentially train on those high-reward states and the probable trajectories leading to them. To this end, we advocate for the use of a \textit{backtracking model} that predicts the preceding states that terminate at a given high-reward state. We can train a model which, starting from a high value state (or one that is estimated to have high value), predicts and samples which (state, action)-tuples may have led to that high value state. These traces of (state, action) pairs, which we refer to as Recall Traces, sampled from this backtracking model starting from a high value state, are informative as they terminate in good states, and hence we can use these traces to improve a policy. We provide a variational interpretation for this idea and a practical algorithm in which the backtracking model samples from an approximate posterior distribution over trajectories which lead to large rewards. Our method improves the sample efficiency of both on- and off-policy RL algorithms across several environments and tasks.

## Fixup Initialization: Residual Learning Without Normalization

- Hongyi Zhang, Yann N. Dauphin, Tengyu Ma
- abstract@open-review(Poster): Normalization layers are a staple in state-of-the-art deep neural network architectures. They are widely believed to stabilize training, enable higher learning rate, accelerate convergence and improve generalization, though the reason for their effectiveness is still an active research topic. In this work, we challenge the commonly-held beliefs by showing that none of the perceived benefits is unique to normalization. Specifically, we propose fixed-update initialization (Fixup), an initialization motivated by solving the exploding and vanishing gradient problem at the

beginning of training via properly rescaling a standard initialization. We find training residual networks with Fixup to be as stable as training with normalization -- even for networks with 10,000 layers. Furthermore, with proper regularization, Fixup enables residual networks without normalization to achieve state-of-the-art performance in image classification and machine translation.

## Diversity-Sensitive Conditional Generative Adversarial Networks

- Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, Honglak Lee
- abstract@open-review(Poster): We propose a simple yet highly effective method that addresses the mode-collapse problem in the Conditional Generative Adversarial Network (cGAN). Although conditional distributions are multi-modal (i.e., having many modes) in practice, most cGAN approaches tend to learn an overly simplified distribution where an input is always mapped to a single output regardless of variations in latent code. To address such issue, we propose to explicitly regularize the generator to produce diverse outputs depending on latent codes. The proposed regularization is simple, general, and can be easily integrated into most conditional GAN objectives. Additionally, explicit regularization on generator allows our method to control a balance between visual quality and diversity. We demonstrate the effectiveness of our method on three conditional generation tasks: image-to-image translation, image inpainting, and future video prediction. We show that simple addition of our regularization to existing models leads to surprisingly diverse generations, substantially outperforming the previous approaches for multi-modal conditional generation specifically designed in each individual task.

## Information asymmetry in KL-regularized RL

- Alexandre Galashov, Siddhant M. Jayakumar, Leonard Hasenclever, Dhruva Tirumala, Jonathan Schwarz, Guillaume Desjardins, Wojciech M. Czarnecki, Yee Whye Teh, Razvan Pascanu, Nicolas Heess
- abstract@open-review(Poster): Many real world tasks exhibit rich structure that is repeated across different parts of the state space or in time. In this work we study the possibility of leveraging such repeated structure to speed up and regularize learning. We start from the KL regularized expected reward objective which introduces an additional component, a default policy. Instead of relying on a fixed default policy, we learn it from data. But crucially, we restrict the amount of information the default policy receives, forcing it to learn reusable behaviors that help the policy learn faster. We formalize this strategy and discuss connections to information bottleneck approaches and to the variational EM algorithm. We present empirical results in both discrete and continuous action domains and demonstrate that, for certain tasks, learning a default policy alongside the policy can significantly speed up and improve learning. Please watch the video demonstrating learned experts and default policies on several continuous control tasks ( https://youtu.be/U2qA3llzus8 ).

## FlowQA: Grasping Flow in History for Conversational Machine Comprehension

- Hsin-Yuan Huang, Eunsol Choi, Wen-tau Yih
- abstract@open-review(Poster): Conversational machine comprehension requires a deep understanding of the conversation history. To enable traditional, single-turn models to encode the history comprehensively, we introduce Flow, a mechanism that can incorporate intermediate representations generated during the process of answering previous questions, through an alternating parallel processing structure. Compared to shallow approaches that concatenate previous questions/answers as input, Flow integrates the latent semantics of the conversation history more deeply. Our model, FlowQA, shows superior performance on two recently proposed conversational challenges (+7.2% F1 on CoQA and +4.0% on QuAC). The effectiveness of Flow also shows in other tasks. By reducing sequential instruction understanding to conversational machine comprehension, FlowQA outperforms the best models on all three domains in SCONE, with +1.8% to +4.4% improvement in accuracy.

## Probabilistic Planning with Sequential Monte Carlo methods

- Alexandre Piche, Valentin Thomas, Cyril Ibrahim, Yoshua Bengio, Chris Pal
- abstract@open-review(Poster): In this work, we propose a novel formulation of planning which views it as a probabilistic inference problem over future optimal trajectories. This enables us to use sampling methods, and thus, tackle planning in continuous domains using a fixed computational budget. We design a new algorithm, Sequential Monte Carlo Planning, by leveraging classical methods in Sequential Monte Carlo and Bayesian smoothing in the context of control as inference. Furthermore, we show that Sequential Monte Carlo Planning can capture multimodal policies and can quickly learn continuous control tasks.

## Spreading vectors for similarity search

- Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Hervé Jégou
- abstract@open-review(Poster): Discretizing floating-point vectors is a fundamental step of modern indexing methods. State-of-the-art techniques learn parameters of the quantizers on training data for optimal performance, thus adapting quantizers to the data. In this work, we propose to reverse this paradigm and adapt the data to the quantizer: we train a neural net whose last layers form a fixed parameter-free quantizer, such as pre-defined points of a sphere. As a proxy objective, we design and train a neural network that favors uniformity in the spherical latent space, while preserving the neighborhood structure after the mapping. For this purpose, we propose a new regularizer derived from the Kozachenko-Leonenko differential entropy estimator and combine it with a locality-aware triplet loss. Experiments show that our end-to-end approach outperforms most learned quantization methods, and is competitive with the state of the art on widely adopted benchmarks. Further more, we show that training without the quantization step results in almost no difference in accuracy, but yields a generic catalyser that can be applied with any subsequent quantization technique.

## LanczosNet: Multi-Scale Deep Graph Convolutional Networks

- Renjie Liao, Zhizhen Zhao, Raquel Urtasun, Richard Zemel
- abstract@open-review(Poster): We propose Lanczos network (LanczosNet) which uses the Lanczos algorithm to construct low rank approximations of the graph Laplacian for graph convolution. Relying on the tridiagonal decomposition of the Lanczos algorithm, we not only efficiently exploit multi-scale information via fast approximated computation of matrix power but also design learnable spectral filters. Being fully differentiable, LanczosNet facilitates both graph kernel learning as well as learning node embeddings. We show the connection between our LanczosNet and graph based manifold learning, especially diffusion maps. We benchmark our model against $8$ recent deep graph networks on citation datasets and QM8 quantum chemistry dataset. Experimental results show that our model achieves the state-of-the-art performance in most tasks.

## Learning what you can do before doing anything

- Oleh Rybkin, Karl Pertsch, Konstantinos G. Derpanis, Kostas Daniilidis, Andrew Jaegle
- abstract@open-review(Poster): Intelligent agents can learn to represent the action spaces of other agents simply by observing them act. Such representations help agents quickly learn to predict the effects of their own actions on the environment and to plan complex action sequences. In this work, we address the problem of learning an agent’s action space purely from visual observation. We use stochastic video prediction to learn a latent variable that captures the scene's dynamics while being minimally sensitive to the scene's static content. We introduce a loss term that encourages the network to capture the composability of visual sequences and show that it leads to representations that disentangle the structure of actions. We call the full model with composable action representations Composable Learned Action Space Predictor (CLASP). We show the applicability of our method to synthetic settings and its potential to capture action spaces in complex, realistic visual settings. When used in a semi-supervised setting, our learned representations perform

comparably to existing fully supervised methods on tasks such as action-conditioned video prediction and planning in the learned action space, while requiring orders of magnitude fewer action labels. Project website: https://daniilidis-group.github.io/learned_action_spaces

## Lagging Inference Networks and Posterior Collapse in Variational Autoencoders

- Junxian He, Daniel Spokoyny, Graham Neubig, Taylor Berg-Kirkpatrick
- abstract@open-review(Poster): The variational autoencoder (VAE) is a popular combination of deep latent variable model and accompanying variational learning technique. By using a neural inference network to approximate the model's posterior on latent variables, VAEs efficiently parameterize a lower bound on marginal data likelihood that can be optimized directly via gradient methods. In practice, however, VAE training often results in a degenerate local optimum known as "posterior collapse" where the model learns to ignore the latent variable and the approximate posterior mimics the prior. In this paper, we investigate posterior collapse from the perspective of training dynamics. We find that during the initial stages of training the inference network fails to approximate the model's true posterior, which is a moving target. As a result, the model is encouraged to ignore the latent encoding and posterior collapse occurs. Based on this observation, we propose an extremely simple modification to VAE training to reduce inference lag: depending on the model's current mutual information between latent variable and observation, we aggressively optimize the inference network before performing each model update. Despite introducing neither new model components nor significant complexity over basic VAE, our approach is able to avoid the problem of collapse that has plagued a large amount of previous work. Empirically, our approach outperforms strong autoregressive baselines on text and image benchmarks in terms of held-out likelihood, and is competitive with more complex techniques for avoiding collapse while being substantially faster.

## Preferences Implicit in the State of the World

- Rohin Shah, Dmitrii Krasheninnikov, Jordan Alexander, Pieter Abbeel, Anca Dragan
- abstract@open-review(Poster): Reinforcement learning (RL) agents optimize only the features specified in a reward function and are indifferent to anything left out inadvertently. This means that we must not only specify what to do, but also the much larger space of what not to do. It is easy to forget these preferences, since these preferences are already satisfied in our environment. This motivates our key insight: when a robot is deployed in an environment that humans act in, the state of the environment is already optimized for what humans want. We can therefore use this implicit preference information from the state to fill in the blanks. We develop an algorithm based on Maximum Causal Entropy IRL and use it to evaluate the idea in a suite of proof-of-concept environments designed to show its properties. We find that information from the initial state can be used to infer both side effects that should be avoided as well as preferences for how the environment should be organized. Our code can be found at https://github.com/HumanCompatibleAI/rlsp.

## A Direct Approach to Robust Deep Learning Using Adversarial Networks

- Huaxia Wang, Chun-Nam Yu
- abstract@open-review(Poster): Deep neural networks have been shown to perform well in many classical machine learning problems, especially in image classification tasks. However, researchers have found that neural networks can be easily fooled, and they are surprisingly sensitive to small perturbations imperceptible to humans. Carefully crafted input images (adversarial examples) can force a well-trained neural network to provide arbitrary outputs. Including adversarial examples during training is a popular defense mechanism against adversarial attacks. In this paper we propose a new defensive mechanism under the generative adversarial network~(GAN) framework. We model the adversarial noise using a generative network, trained jointly with a classification discriminative network as a minimax game. We show empirically that our adversarial network approach works well against black box attacks, with performance on par with state-of-art methods such as ensemble adversarial training and adversarial training with projected gradient descent.

## Improving the Generalization of Adversarial Training with Domain Adaptation

- Chuanbiao Song, Kun He, Liwei Wang, John E. Hopcroft
- abstract@open-review(Poster): By injecting adversarial examples into training data, adversarial training is promising for improving the robustness of deep learning models. However, most existing adversarial training approaches are based on a specific type of adversarial attack. It may not provide sufficiently representative samples from the adversarial domain, leading to a weak generalization ability on adversarial examples from other attacks. Moreover, during the adversarial training, adversarial perturbations on inputs are usually crafted by fast single-step adversaries so as to scale to large datasets. This work is mainly focused on the adversarial training yet efficient FGSM adversary. In this scenario, it is difficult to train a model with great generalization due to the lack of representative adversarial samples, aka the samples are unable to accurately reflect the adversarial domain. To alleviate this problem, we propose a novel Adversarial Training with Domain Adaptation (ATDA) method. Our intuition is to regard the adversarial training on FGSM adversary as a domain adaption task with limited number of target domain samples. The main idea is to learn a representation that is semantically meaningful and domain invariant on the clean domain as well as the adversarial domain. Empirical evaluations on Fashion-MNIST, SVHN, CIFAR-10 and CIFAR-100 demonstrate that ATDA can greatly improve the generalization of adversarial training and the smoothness of the learned models, and outperforms state-of-the-art methods on standard benchmark datasets. To show the transfer ability of our method, we also extend ATDA to the adversarial training on iterative attacks such as PGD-Adversial Training (PAT) and the defense performance is improved considerably.

## Subgradient Descent Learns Orthogonal Dictionaries

- Yu Bai, Qijia Jiang, Ju Sun
- abstract@open-review(Poster): This paper concerns dictionary learning, i.e., sparse coding, a fundamental representation learning problem. We show that a subgradient descent algorithm, with random initialization, can recover orthogonal dictionaries on a natural nonsmooth, nonconvex L1 minimization formulation of the problem, under mild statistical assumption on the data. This is in contrast to previous provable methods that require either expensive computation or delicate initialization schemes. Our analysis develops several tools for characterizing landscapes of nonsmooth functions, which might be of independent interest for provable training of deep networks with nonsmooth activations (e.g., ReLU), among other applications. Preliminary synthetic and real experiments corroborate our analysis and show that our algorithm works well empirically in recovering orthogonal dictionaries.

## Improving Differentiable Neural Computers Through Memory Masking, De-allocation, and Link Distribution Sharpness Control

- Robert Csordas, Juergen Schmidhuber
- abstract@open-review(Poster): The Differentiable Neural Computer (DNC) can learn algorithmic and question answering tasks. An analysis of its internal activation patterns reveals three problems: Most importantly, the lack of key-value separation makes the address distribution resulting from content-based look-up noisy and flat, since the value influences the score calculation, although only the key should. Second, DNC's de-allocation of memory results in aliasing, which is a problem for content-based look-up. Thirdly, chaining memory reads with the temporal linkage matrix exponentially degrades the quality of the address distribution. Our proposed fixes of these problems yield improved performance on arithmetic tasks, and also improve the mean error rate on the bAbI question answering dataset by 43%.

## Unsupervised Control Through Non-Parametric Discriminative Rewards

- David Warde-Farley, Tom Van de Wiele, Tejas Kulkarni, Catalin Ionescu, Steven Hansen, Volodymyr Mnih

- abstract@open-review(Poster): Learning to control an environment without hand-crafted rewards or expert data remains challenging and is at the frontier of reinforcement learning research. We present an unsupervised learning algorithm to train agents to achieve perceptually-specified goals using only a stream of observations and actions. Our agent simultaneously learns a goal-conditioned policy and a goal achievement reward function that measures how similar a state is to the goal state. This dual optimization leads to a co-operative game, giving rise to a learned reward function that reflects similarity in controllable aspects of the environment instead of distance in the space of observations. We demonstrate the efficacy of our agent to learn, in an unsupervised manner, to reach a diverse set of goals on three domains -- Atari, the DeepMind Control Suite and DeepMind Lab.

## Self-Tuning Networks: Bilevel Optimization of Hyperparameters using Structured Best-Response Functions

- Matthew Mackay, Paul Vicol, Jonathan Lorraine, David Duvenaud, Roger Grosse
- abstract@open-review(Poster): Hyperparameter optimization can be formulated as a bilevel optimization problem, where the optimal parameters on the training set depend on the hyperparameters. We aim to adapt regularization hyperparameters for neural networks by fitting compact approximations to the best-response function, which maps hyperparameters to optimal weights and biases. We show how to construct scalable best-response approximations for neural networks by modeling the best-response as a single network whose hidden units are gated conditionally on the regularizer. We justify this approximation by showing the exact best-response for a shallow linear network with L2-regularized Jacobian can be represented by a similar gating mechanism. We fit this model using a gradient-based hyperparameter optimization algorithm which alternates between approximating the best-response around the current hyperparameters and optimizing the hyperparameters using the approximate best-response function. Unlike other gradient-based approaches, we do not require differentiating the training loss with respect to the hyperparameters, allowing us to tune discrete hyperparameters, data augmentation hyperparameters, and dropout probabilities. Because the hyperparameters are adapted online, our approach discovers hyperparameter schedules that can outperform fixed hyperparameter values. Empirically, our approach outperforms competing hyperparameter optimization methods on large-scale deep learning problems. We call our networks, which update their own hyperparameters online during training, Self-Tuning Networks (STNs).

## Explaining Image Classifiers by Counterfactual Generation

- Chun-Hao Chang, Elliot Creager, Anna Goldenberg, David Duvenaud
- abstract@open-review(Poster): When an image classifier makes a prediction, which parts of the image are relevant and why? We can rephrase this question to ask: which parts of the image, if they were not seen by the classifier, would most change its decision? Producing an answer requires marginalizing over images that could have been seen but weren't. We can sample plausible image in-fills by conditioning a generative model on the rest of the image. We then optimize to find the image regions that most change the classifier's decision after in-fill. Our approach contrasts with ad-hoc in-filling approaches, such as blurring or injecting noise, which generate inputs far from the data distribution, and ignore informative relationships between different parts of the image. Our method produces more compact and relevant saliency maps, with fewer artifacts compared to previous methods.

## Predicting the Generalization Gap in Deep Networks with Margin Distributions

- Yiding Jiang, Dilip Krishnan, Hossein Mobahi, Samy Bengio
- abstract@open-review(Poster): As shown in recent research, deep neural networks can perfectly fit randomly labeled data, but with very poor accuracy on held out data. This phenomenon indicates that loss functions such as cross-entropy are not a reliable indicator of generalization. This leads to the crucial question of how generalization gap should be predicted from the training data and network parameters. In this paper, we propose such a measure, and conduct extensive empirical studies on how well it can predict the generalization gap. Our measure is based on the concept of margin distribution, which are the distances of training points to the decision boundary. We find that it is necessary to use margin distributions at multiple layers of a deep network. On the CIFAR-10 and the CIFAR-100 datasets, our proposed measure correlates very strongly with the generalization gap. In addition, we find the following other factors to be of importance: normalizing margin values for scale independence, using characterizations of margin distribution rather than just the margin (closest distance to decision boundary), and working in log space instead of linear space (effectively using a product of margins rather than a sum). Our measure can be easily applied to feedforward deep networks with any architecture and may point towards new training loss functions that could enable better generalization.

## Mode Normalization

- Lucas Deecke, Iain Murray, Hakan Bilen
- abstract@open-review(Poster): Normalization methods are a central building block in the deep learning toolbox. They accelerate and stabilize training, while decreasing the dependence on manually tuned learning rate schedules. When learning from multi-modal distributions, the effectiveness of batch normalization (BN), arguably the most prominent normalization method, is reduced. As a remedy, we propose a more flexible approach: by extending the normalization to more than a single mean and variance, we detect modes of data on-the-fly, jointly normalizing samples that share common features. We demonstrate that our method outperforms BN and other widely used normalization techniques in several experiments, including single and multi-task datasets.

## Kernel Change-point Detection with Auxiliary Deep Generative Models

- Wei-Cheng Chang, Chun-Liang Li, Yiming Yang, Barnabás Póczos
- abstract@open-review(Poster): Detecting the emergence of abrupt property changes in time series is a challenging problem. Kernel two-sample test has been studied for this task which makes fewer assumptions on the distributions than traditional parametric approaches. However, selecting kernels is non-trivial in practice. Although kernel selection for the two-sample test has been studied, the insufficient samples in change point detection problem hinder the success of those developed kernel selection algorithms. In this paper, we propose KL-CPD, a novel kernel learning framework for time series CPD that optimizes a lower bound of test power via an auxiliary generative model. With deep kernel parameterization, KL-CPD endows kernel two-sample test with the data-driven kernel to detect different types of change-points in real-world applications. The proposed approach significantly outperformed other state-of-the-art methods in our comparative evaluation of benchmark datasets and simulation studies.

## Towards Metamerism via Foveated Style Transfer

- Arturo Deza, Aditya Jonnalagadda, Miguel P. Eckstein
- abstract@open-review(Poster): The problem of visual metamerism is defined as finding a family of perceptually indistinguishable, yet physically different images. In this paper, we propose our NeuroFovea metamer model, a foveated generative model that is based on a mixture of peripheral representations and style transfer forward-pass algorithms. Our gradient-descent free model is parametrized by a foveated VGG19 encoder-decoder which allows us to encode images in high dimensional space and interpolate between the content and texture information with adaptive instance normalization anywhere in the visual field. Our contributions include: 1) A framework for computing metamers that resembles a noisy communication system via a foveated feed-forward encoder-decoder network – We observe that metamerism arises as a byproduct of noisy perturbations that partially lie in the perceptual null space; 2) A perceptual optimization scheme as a solution to the hyperparametric nature of our metamer model that requires tuning of the image-texture tradeoff coefficients everywhere in the visual field which are a consequence of internal noise; 3) An ABX psychophysical evaluation of our metamers where we also find that the rate of growth of the receptive fields in our model match V1 for reference metamers and V2 between synthesized samples. Our model also renders metamers at roughly a second, presenting a ×1000 speed-up compared to the previous work, which now allows for tractable data-driven metamer experiments.

## Learning To Solve Circuit-SAT: An Unsupervised Differentiable Approach

- Saeed Amizadeh, Sergiy Matusevych, Markus Weimer
- abstract@open-review(Poster): Recent efforts to combine Representation Learning with Formal Methods, commonly known as the Neuro-Symbolic Methods, have given rise to a new trend of applying rich neural architectures to solve classical combinatorial optimization problems. In this paper, we propose a neural framework that can learn to solve the Circuit Satisfiability problem. Our framework is built upon two fundamental contributions: a rich embedding architecture that encodes the problem structure and an end-to-end differentiable training procedure that mimics Reinforcement Learning and trains the model directly toward solving the SAT problem. The experimental results show the superior out-of-sample generalization performance of our framework compared to the recently developed NeuroSAT method.

## Variance Reduction for Reinforcement Learning in Input-Driven Environments

- Hongzi Mao, Shaileshh Bojja Venkatakrishnan, Malte Schwarzkopf, Mohammad Alizadeh
- abstract@open-review(Poster): We consider reinforcement learning in input-driven environments, where an exogenous, stochastic input process affects the dynamics of the system. Input processes arise in many applications, including queuing systems, robotics control with disturbances, and object tracking. Since the state dynamics and rewards depend on the input process, the state alone provides limited information for the expected future returns. Therefore, policy gradient methods with standard state-dependent baselines suffer high variance during training. We derive a bias-free, input-dependent baseline to reduce this variance, and analytically show its benefits over state-dependent baselines. We then propose a meta-learning approach to overcome the complexity of learning a baseline that depends on a long sequence of inputs. Our experimental results show that across environments from queuing systems, computer networks, and MuJoCo robotic locomotion, input-dependent baselines consistently improve training stability and result in better eventual policies.

## Robustness May Be at Odds with Accuracy

- Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, Aleksander Madry
- abstract@open-review(Poster): We show that there exists an inherent tension between the goal of adversarial robustness and that of standard generalization. Specifically, training robust models may not only be more resource-consuming, but also lead to a reduction of standard accuracy. We demonstrate that this trade-off between the standard accuracy of a model and its robustness to adversarial perturbations provably exists even in a fairly simple and natural setting. These findings also corroborate a similar phenomenon observed in practice. Further, we argue that this phenomenon is a consequence of robust classifiers learning fundamentally different feature representations than standard classifiers. These differences, in particular, seem to result in unexpected benefits: the features learned by robust models tend to align better with salient data characteristics and human perception.

## Supervised Community Detection with Line Graph Neural Networks

- Zhengdao Chen, Lisha Li, Joan Bruna
- abstract@open-review(Poster): Community detection in graphs can be solved via spectral methods or posterior inference under certain probabilistic graphical models. Focusing on random graph families such as the stochastic block model, recent research has unified both approaches and identified both statistical and computational detection thresholds in terms of the signal-to-noise ratio. By recasting community detection as a node-wise classification problem on graphs, we can also study it from a learning perspective. We present a novel family of Graph Neural Networks (GNNs) for solving community detection problems in a supervised learning setting. We show that, in a data-driven manner and without access to the underlying generative models, they can match or even surpass the performance of the belief propagation algorithm on binary and multiclass stochastic block models, which is believed to reach the computational threshold in these cases. In particular, we propose to augment GNNs with the non-backtracking operator defined on the line graph of edge adjacencies. The GNNs are achieved good performance on real-world datasets. In addition, we perform the first analysis of the optimization landscape of using (linear) GNNs to solve community detection problems, demonstrating that under certain simplifications and assumptions, the loss value at any local minimum is close to the loss value at the global minimum/minima.

## BA-Net: Dense Bundle Adjustment Networks

- Chengzhou Tang, Ping Tan
- abstract@open-review(Oral): This paper introduces a network architecture to solve the structure-from-motion (SfM) problem via feature-metric bundle adjustment (BA), which explicitly enforces multi-view geometry constraints in the form of feature-metric error. The whole pipeline is differentiable, so that the network can learn suitable features that make the BA problem more tractable. Furthermore, this work introduces a novel depth parameterization to recover dense per-pixel depth. The network first generates several basis depth maps according to the input image, and optimizes the final depth as a linear combination of these basis depth maps via feature-metric BA. The basis depth maps generator is also learned via end-to-end training. The whole system nicely combines domain knowledge (i.e. hard-coded multi-view geometry constraints) and deep learning (i.e. feature learning and basis depth maps learning) to address the challenging dense SfM problem. Experiments on large scale real data prove the success of the proposed method.

## Harmonic Unpaired Image-to-image Translation

- Rui Zhang, Tomas Pfister, Jia Li
- abstract@open-review(Poster): The recent direction of unpaired image-to-image translation is on one hand very exciting as it alleviates the big burden in obtaining label-intensive pixel-to-pixel supervision, but it is on the other hand not fully satisfactory due to the presence of artifacts and degenerated transformations. In this paper, we take a manifold view of the problem by introducing a smoothness term over the sample graph to attain harmonic functions to enforce consistent mappings during the translation. We develop HarmonicGAN to learn bi-directional translations between the source and the target domains. With the help of similarity-consistency, the inherent self-consistency property of samples can be maintained. Distance metrics defined on two types of features including histogram and CNN are exploited. Under an identical problem setting as CycleGAN, without additional manual inputs and only at a small training-time cost, HarmonicGAN demonstrates a significant qualitative and quantitative improvement over the state of the art, as well as improved interpretability. We show experimental results in a number of applications including medical imaging, object transfiguration, and semantic labeling. We outperform the competing methods in all tasks, and for a medical imaging task in particular our method turns CycleGAN from a failure to a success, halving the mean-squared error, and generating images that radiologists prefer over competing methods in 95% of cases.

## Learning Neural PDE Solvers with Convergence Guarantees

- Jun-Ting Hsieh, Shengjia Zhao, Stephan Eismann, Lucia Mirabella, Stefano Ermon
- abstract@open-review(Poster): Partial differential equations (PDEs) are widely used across the physical and computational sciences. Decades of research and engineering went into designing fast iterative solution methods. Existing solvers are general purpose, but may be sub-optimal for specific classes of problems. In contrast to existing hand-crafted solutions, we propose an approach to learn a fast iterative solver tailored to a specific domain. We achieve this goal by learning to modify the updates of an existing solver using a deep neural network. Crucially, our approach is proven to preserve strong correctness and convergence guarantees. After training on a single geometry, our model generalizes to a wide variety of geometries and boundary conditions, and achieves 2-3 times speedup compared to state-of-the-art solvers.

## Meta-learning with differentiable closed-form solvers

- Luca Bertinetto, Joao F. Henriques, Philip Torr, Andrea Vedaldi
- abstract@open-review(Poster): Adapting deep networks to new concepts from a few examples is challenging, due to the high computational requirements of standard fine-tuning procedures. Most work on few-shot learning has thus focused on simple learning techniques for adaptation, such as nearest neighbours or gradient descent. Nonetheless, the machine learning literature contains a wealth of methods that learn non-deep models very efficiently. In this paper, we propose to use these fast convergent methods as the main adaptation mechanism for few-shot learning. The main idea is to teach a deep network to use standard machine learning tools, such as ridge regression, as part of its own internal model, enabling it to quickly adapt to novel data. This requires back-propagating errors through the solver steps. While normally the cost of the matrix operations involved in such a process would be significant, by using the Woodbury identity we can make the small number of examples work to our advantage. We propose both closed-form and iterative solvers, based on ridge regression and logistic regression components. Our methods constitute a simple and novel approach to the problem of few-shot learning and achieve performance competitive with or superior to the state of the art on three benchmarks.

## Building Dynamic Knowledge Graphs from Text using Machine Reading Comprehension

- Rajarshi Das, Tsendsuren Munkhdalai, Xingdi Yuan, Adam Trischler, Andrew McCallum
- abstract@open-review(Poster): We propose a neural machine-reading model that constructs dynamic knowledge graphs from procedural text. It builds these graphs recurrently for each step of the described procedure, and uses them to track the evolving states of participant entities. We harness and extend a recently proposed machine reading comprehension(MRC) model to query for entity states, since these states are generally communicated in spans of text and MRC models perform well in extracting entity-centric spans. The explicit, structured, and evolving knowledge graph representations that our model constructs can be used in downstream question answering tasks to improve machine comprehension of text, as we demonstrate empirically. On two comprehension tasks from the recently proposed ProPara dataset, our model achieves state-of-the-art results. We further show that our model is competitive on the Recipes dataset, suggesting it may be generally applicable.

## Prior Convictions: Black-box Adversarial Attacks with Bandits and Priors

- Andrew Ilyas, Logan Engstrom, Aleksander Madry
- abstract@open-review(Poster): We study the problem of generating adversarial examples in a black-box setting in which only loss-oracle access to a model is available. We introduce a framework that conceptually unifies much of the existing work on black-box attacks, and demonstrate that the current state-of-the-art methods are optimal in a natural sense. Despite this optimality, we show how to improve black-box attacks by bringing a new element into the problem: gradient priors. We give a bandit optimization-based algorithm that allows us to seamlessly integrate any such priors, and we explicitly identify and incorporate two examples. The resulting methods use two to four times fewer queries and fail two to five times less than the current state-of-the-art. The code for reproducing our work is available at https://git.io/fAjOJ.

## DHER: Hindsight Experience Replay for Dynamic Goals

- Meng Fang, Cheng Zhou, Bei Shi, Boqing Gong, Jia Xu, Tong Zhang
- abstract@open-review(Poster): Dealing with sparse rewards is one of the most important challenges in reinforcement learning (RL), especially when a goal is dynamic (e.g., to grasp a moving object). Hindsight experience replay (HER) has been shown an effective solution to handling sparse rewards with fixed goals. However, it does not account for dynamic goals in its vanilla form and, as a result, even degrades the performance of existing off-policy RL algorithms when the goal is changing over time.

In this paper, we present Dynamic Hindsight Experience Replay (DHER), a novel approach for tasks with dynamic goals in the presence of sparse rewards. DHER automatically assembles successful experiences from two relevant failures and can be used to enhance an arbitrary off-policy RL algorithm when the tasks' goals are dynamic. We evaluate DHER on tasks of robotic manipulation and moving object tracking, and transfer the polices from simulation to physical robots. Extensive comparison and ablation studies demonstrate the superiority of our approach, showing that DHER is a crucial ingredient to enable RL to solve tasks with dynamic goals in manipulation and grid world domains.

## Stochastic Gradient/Mirror Descent: Minimax Optimality and Implicit Regularization

- Navid Azizan, Babak Hassibi
- abstract@open-review(Poster): Stochastic descent methods (of the gradient and mirror varieties) have become increasingly popular in optimization. In fact, it is now widely recognized that the success of deep learning is not only due to the special deep architecture of the models, but also due to the behavior of the stochastic descent methods used, which play a key role in reaching "good" solutions that generalize well to unseen data. In an attempt to shed some light on why this is the case, we revisit some minimax properties of stochastic gradient descent (SGD) for the square loss of linear models---originally developed in the 1990's---and extend them to \emph{general} stochastic mirror descent (SMD) algorithms for \emph{general} loss functions and \emph{nonlinear} models. In particular, we show that there is a fundamental identity which holds for SMD (and SGD) under very general conditions, and which implies the minimax optimality of SMD (and SGD) for sufficiently small step size, and for a general class of loss functions and general nonlinear models. We further show that this identity can be used to naturally establish other properties of SMD (and SGD), namely convergence and \emph{implicit regularization} for over-parameterized linear models (in what is now being called the "interpolating regime"), some of which have been shown in certain cases in prior literature. We also argue how this identity can be used in the so-called "highly over-parameterized" nonlinear setting (where the number of parameters far exceeds the number of data points) to provide insights into why SMD (and SGD) may have similar convergence and implicit regularization properties for deep learning.

## Unsupervised Learning of the Set of Local Maxima

- Lior Wolf, Sagie Benaim, Tomer Galanti
- abstract@open-review(Poster): This paper describes a new form of unsupervised learning, whose input is a set of unlabeled points that are assumed to be local maxima of an unknown value function $v$ in an unknown subset of the vector space. Two functions are learned: (i) a set indicator $c$, which is a binary classifier, and (ii) a comparator function $h$ that given two nearby samples, predicts which sample has the higher value of the unknown function $v$. Loss terms are used to ensure that all training samples $\vx$ are a local maxima of $v$, according to $h$ and satisfy $c(\vx)=1$. Therefore, $c$ and $h$ provide training signals to each other: a point $\vx'$ in the vicinity of $\vx$ satisfies $c(\vx)=-1$ or is deemed by $h$ to be lower in value than $\vx$. We present an algorithm, show an example where it is more efficient to use local maxima as an indicator function than to employ conventional classification, and derive a suitable generalization bound. Our experiments show that the method is able to outperform one-class classification algorithms in the task of anomaly detection and also provide an additional signal that is extracted in a completely unsupervised way.

## Neural Logic Machines

- Honghua Dong, Jiayuan Mao, Tian Lin, Chong Wang, Lihong Li, Denny Zhou
- abstract@open-review(Poster): We propose the Neural Logic Machine (NLM), a neural-symbolic architecture for both inductive learning and logic reasoning. NLMs exploit the power of both neural networks---as function approximators, and logic programming---as a symbolic processor for objects with properties, relations, logic connectives, and quantifiers. After being trained on small-scale tasks (such as sorting short arrays), NLMs can recover

lifted rules, and generalize to large-scale tasks (such as sorting longer arrays). In our experiments, NLMs achieve perfect generalization in a number of tasks, from relational reasoning tasks on the family tree and general graphs, to decision making tasks including sorting arrays, finding shortest paths, and playing the blocks world. Most of these tasks are hard to accomplish for neural networks or inductive logic programming alone.

## Defensive Quantization: When Efficiency Meets Robustness

- Ji Lin, Chuang Gan, Song Han
- abstract@open-review(Poster): Neural network quantization is becoming an industry standard to efficiently deploy deep learning models on hardware platforms, such as CPU, GPU, TPU, and FPGAs. However, we observe that the conventional quantization approaches are vulnerable to adversarial attacks. This paper aims to raise people's awareness about the security of the quantized models, and we designed a novel quantization methodology to jointly optimize the efficiency and robustness of deep learning models. We first conduct an empirical study to show that vanilla quantization suffers more from adversarial attacks. We observe that the inferior robustness comes from the error amplification effect, where the quantization operation further enlarges the distance caused by amplified noise. Then we propose a novel Defensive Quantization (DQ) method by controlling the Lipschitz constant of the network during quantization, such that the magnitude of the adversarial noise remains non-expansive during inference. Extensive experiments on CIFAR-10 and SVHN datasets demonstrate that our new quantization method can defend neural networks against adversarial examples, and even achieves superior robustness than their full-precision counterparts, while maintaining the same hardware efficiency as vanilla quantization approaches. As a by-product, DQ can also improve the accuracy of quantized models without adversarial attack.

## Dimensionality Reduction for Representing the Knowledge of Probabilistic Models

- Marc T Law, Jake Snell, Amir-massoud Farahmand, Raquel Urtasun, Richard S Zemel
- abstract@open-review(Poster): Most deep learning models rely on expressive high-dimensional representations to achieve good performance on tasks such as classification. However, the high dimensionality of these representations makes them difficult to interpret and prone to over-fitting. We propose a simple, intuitive and scalable dimension reduction framework that takes into account the soft probabilistic interpretation of standard deep models for classification. When applying our framework to visualization, our representations more accurately reflect inter-class distances than standard visualization techniques such as t-SNE. We show experimentally that our framework improves generalization performance to unseen categories in zero-shot learning. We also provide a finite sample error upper bound guarantee for the method.

## Biologically-Plausible Learning Algorithms Can Scale to Large Datasets

- Will Xiao, Honglin Chen, Qianli Liao, Tomaso Poggio
- abstract@open-review(Poster): The backpropagation (BP) algorithm is often thought to be biologically implausible in the brain. One of the main reasons is that BP requires symmetric weight matrices in the feedforward and feedback pathways. To address this “weight transport problem” (Grossberg, 1987), two biologically-plausible algorithms, proposed by Liao et al. (2016) and Lillicrap et al. (2016), relax BP’s weight symmetry requirements and demonstrate comparable learning capabilities to that of BP on small datasets. However, a recent study by Bartunov et al. (2018) finds that although feedback alignment (FA) and some variants of target-propagation (TP) perform well on MNIST and CIFAR, they perform significantly worse than BP on ImageNet. Here, we additionally evaluate the sign-symmetry (SS) algorithm (Liao et al., 2016), which differs from both BP and FA in that the feedback and feedforward weights do not share magnitudes but share signs. We examined the performance of sign-symmetry and feedback alignment on ImageNet and MS COCO datasets using different network architectures (ResNet-18 and AlexNet for ImageNet; RetinaNet for MS COCO). Surprisingly, networks trained with sign-symmetry can attain classification performance approaching that of BP-trained networks. These results complement the study by Bartunov et al. (2018) and establish a new benchmark for future biologically-plausible learning algorithms on more difficult datasets and more complex architectures.

## Generating Multi-Agent Trajectories using Programmatic Weak Supervision

- Eric Zhan, Stephan Zheng, Yisong Yue, Long Sha, Patrick Lucey
- abstract@open-review(Poster): We study the problem of training sequential generative models for capturing coordinated multi-agent trajectory behavior, such as offensive basketball gameplay. When modeling such settings, it is often beneficial to design hierarchical models that can capture long-term coordination using intermediate variables. Furthermore, these intermediate variables should capture interesting high-level behavioral semantics in an interpretable and manipulable way. We present a hierarchical framework that can effectively learn such sequential generative models. Our approach is inspired by recent work on leveraging programmatically produced weak labels, which we extend to the spatiotemporal regime. In addition to synthetic settings, we show how to instantiate our framework to effectively model complex interactions between basketball players and generate realistic multi-agent trajectories of basketball gameplay over long time periods. We validate our approach using both quantitative and qualitative evaluations, including a user study comparison conducted with professional sports analysts.

## Predict then Propagate: Graph Neural Networks meet Personalized PageRank

- Johannes Gasteiger, Aleksandar Bojchevski, Stephan Günnemann
- abstract@open-review(Poster): Neural message passing algorithms for semi-supervised classification on graphs have recently achieved great success. However, for classifying a node these methods only consider nodes that are a few propagation steps away and the size of this utilized neighborhood is hard to extend. In this paper, we use the relationship between graph convolutional networks (GCN) and PageRank to derive an improved propagation scheme based on personalized PageRank. We utilize this propagation procedure to construct a simple model, personalized propagation of neural predictions (PPNP), and its fast approximation, APPNP. Our model's training time is on par or faster and its number of parameters on par or lower than previous models. It leverages a large, adjustable neighborhood for classification and can be easily combined with any neural network. We show that this model outperforms several recently proposed methods for semi-supervised classification in the most thorough study done so far for GCN-like models. Our implementation is available online.

## Environment Probing Interaction Policies

- Wenxuan Zhou, Lerrel Pinto, Abhinav Gupta
- abstract@open-review(Poster): A key challenge in reinforcement learning (RL) is environment generalization: a policy trained to solve a task in one environment often fails to solve the same task in a slightly different test environment. A common approach to improve inter-environment transfer is to learn policies that are invariant to the distribution of testing environments. However, we argue that instead of being invariant, the policy should identify the specific nuances of an environment and exploit them to achieve better performance. In this work, we propose the “Environment-Probing” Interaction (EPI) policy, a policy that probes a new environment to extract an implicit understanding of that environment’s behavior. Once this environment-specific information is obtained, it is used as an additional input to a task-specific policy that can now perform environment-conditioned actions to solve a task. To learn these EPI-policies, we present a reward function based on transition predictability. Specifically, a higher reward is given if the trajectory generated by the EPI-policy can be used to better predict transitions. We experimentally show that EPI-conditioned task-specific policies significantly outperform commonly used policy generalization methods on novel testing environments.

## Transferring Knowledge across Learning Processes

- Sebastian Flennerhag, Pablo G. Moreno, Neil D. Lawrence, Andreas Damianou
- abstract@open-review(Oral): In complex transfer learning scenarios new tasks might not be tightly linked to previous tasks. Approaches that transfer information contained only in the final parameters of a source model will therefore struggle. Instead, transfer learning at at higher level of abstraction is needed. We propose Leap, a framework that achieves this by transferring knowledge across learning processes. We associate each task with a manifold on which the training process travels from initialization to final parameters and construct a meta-learning objective that minimizes the expected length of this path. Our framework leverages only information obtained during training and can be computed on the fly at negligible cost. We demonstrate that our framework outperforms competing methods, both in meta-learning and transfer learning, on a set of computer vision tasks. Finally, we demonstrate that Leap can transfer knowledge across learning processes in demanding reinforcement learning environments (Atari) that involve millions of gradient steps.

## Time-Agnostic Prediction: Predicting Predictable Video Frames

- Dinesh Jayaraman, Frederik Ebert, Alexei Efros, Sergey Levine
- abstract@open-review(Poster): Prediction is arguably one of the most basic functions of an intelligent system. In general, the problem of predicting events in the future or between two waypoints is exceedingly difficult. However, most phenomena naturally pass through relatively predictable bottlenecks---while we cannot predict the precise trajectory of a robot arm between being at rest and holding an object up, we can be certain that it must have picked the object up. To exploit this, we decouple visual prediction from a rigid notion of time. While conventional approaches predict frames at regularly spaced temporal intervals, our time-agnostic predictors (TAP) are not tied to specific times so that they may instead discover predictable "bottleneck" frames no matter when they occur. We evaluate our approach for future and intermediate frame prediction across three robotic manipulation tasks. Our predictions are not only of higher visual quality, but also correspond to coherent semantic subgoals in temporally extended tasks.

## Unsupervised Adversarial Image Reconstruction

- Arthur Pajot, Emmanuel de Bezenac, Patrick Gallinari
- abstract@open-review(Poster): We address the problem of recovering an underlying signal from lossy, inaccurate observations in an unsupervised setting. Typically, we consider situations where there is little to no background knowledge on the structure of the underlying signal, no access to signal-measurement pairs, nor even unpaired signal-measurement data. The only available information is provided by the observations and the measurement process statistics. We cast the problem as finding the \textit{maximum a posteriori} estimate of the signal given each measurement, and propose a general framework for the reconstruction problem. We use a formulation of generative adversarial networks, where the generator takes as input a corrupted observation in order to produce realistic reconstructions, and add a penalty term tying the reconstruction to the associated observation. We evaluate our reconstructions on several image datasets with different types of corruptions. The proposed approach yields better results than alternative baselines, and comparable performance with model variants trained with additional supervision.

## Deep Decoder: Concise Image Representations from Untrained Non-convolutional Networks

- Reinhard Heckel, Paul Hand
- abstract@open-review(Poster): Deep neural networks, in particular convolutional neural networks, have become highly effective tools for compressing images and solving inverse problems including denoising, inpainting, and reconstruction from few and noisy measurements. This success can be attributed in part to their ability to represent and generate natural images well. Contrary to classical tools such as wavelets, image-generating deep neural networks have a large number of parameters---typically a multiple of their output dimension---and need to be trained on large datasets. In this paper, we propose an untrained simple image model, called the deep decoder, which is a deep neural network that can generate natural images from very few weight parameters. The deep decoder has a simple architecture with no convolutions and fewer weight parameters than the output dimensionality. This underparameterization enables the deep decoder to compress images into a concise set of network weights, which we show is on par with wavelet-based thresholding. Further, underparameterization provides a barrier to overfitting, allowing the deep decoder to have state-of-the-art performance for denoising. The deep decoder is simple in the sense that each layer has an identical structure that consists of only one upsampling unit, pixel-wise linear combination of channels, ReLU activation, and channelwise normalization. This simplicity makes the network amenable to theoretical analysis, and it sheds light on the aspects of neural networks that enable them to form effective signal representations.

## Minimal Images in Deep Neural Networks: Fragile Object Recognition in Natural Images

- Sanjana Srivastava, Guy Ben-Yosef, Xavier Boix
- abstract@open-review(Poster): The human ability to recognize objects is impaired when the object is not shown in full. "Minimal images" are the smallest regions of an image that remain recognizable for humans. Ullman et al. (2016) show that a slight modification of the location and size of the visible region of the minimal image produces a sharp drop in human recognition accuracy. In this paper, we demonstrate that such drops in accuracy due to changes of the visible region are a common phenomenon between humans and existing state-of-the-art deep neural networks (DNNs), and are much more prominent in DNNs. We found many cases where DNNs classified one region correctly and the other incorrectly, though they only differed by one row or column of pixels, and were often bigger than the average human minimal image size. We show that this phenomenon is independent from previous works that have reported lack of invariance to minor modifications in object location in DNNs. Our results thus reveal a new failure mode of DNNs that also affects humans to a much lesser degree. They expose how fragile DNN recognition ability is in natural images even without adversarial patterns being introduced. Bringing the robustness of DNNs in natural images to the human level remains an open challenge for the community.

## Overcoming the Disentanglement vs Reconstruction Trade-off via Jacobian Supervision

- José Lezama
- abstract@open-review(Poster): A major challenge in learning image representations is the disentangling of the factors of variation underlying the image formation. This is typically achieved with an autoencoder architecture where a subset of the latent variables is constrained to correspond to specific factors, and the rest of them are considered nuisance variables. This approach has an important drawback: as the dimension of the nuisance variables is increased, image reconstruction is improved, but the decoder has the flexibility to ignore the specified factors, thus losing the ability to condition the output on them. In this work, we propose to overcome this trade-off by progressively growing the dimension of the latent code, while constraining the Jacobian of the output image with respect to the disentangled variables to remain the same. As a result, the obtained models are effective at both disentangling and reconstruction. We demonstrate the applicability of this method in both unsupervised and supervised scenarios for learning disentangled representations. In a facial attribute manipulation task, we obtain high quality image generation while smoothly controlling dozens of attributes with a single model. This is an order of magnitude more disentangled factors than state-of-the-art methods, while obtaining visually similar or superior results, and avoiding adversarial training.

## ProbGAN: Towards Probabilistic GAN with Theoretical Guarantees

- Hao He, Hao Wang, Guang-He Lee, Yonglong Tian
- abstract@open-review(Poster): Probabilistic modelling is a principled framework to perform model aggregation, which has been a primary mechanism to combat mode collapse in the context of Generative Adversarial Networks (GAN). In this paper, we propose a novel probabilistic framework for GANs, ProbGAN, which iteratively learns a distribution over generators with a carefully crafted prior. Learning is efficiently triggered by a tailored stochastic gradient Hamiltonian Monte Carlo with a novel gradient approximation to perform Bayesian inference. Our theoretical analysis further reveals that our treatment is the first probabilistic framework that yields an equilibrium where generator distributions are faithful to the data distribution. Empirical

evidence on synthetic high-dimensional multi-modal data and image databases (CIFAR-10, STL-10, and ImageNet) demonstrates the superiority of our method over both start-of-the-art multi-generator GANs and other probabilistic treatment for GANs.

## Theoretical Analysis of Auto Rate-Tuning by Batch Normalization

- Sanjeev Arora, Zhiyuan Li, Kaifeng Lyu
- abstract@open-review(Poster): Batch Normalization (BN) has become a cornerstone of deep learning across diverse architectures, appearing to help optimization as well as generalization. While the idea makes intuitive sense, theoretical analysis of its effectiveness has been lacking. Here theoretical support is provided for one of its conjectured properties, namely, the ability to allow gradient descent to succeed with less tuning of learning rates. It is shown that even if we fix the learning rate of scale-invariant parameters (e.g., weights of each layer with BN) to a constant (say, 0.3), gradient descent still approaches a stationary point (i.e., a solution where gradient is zero) in the rate of T^{−1/2} in T iterations, asymptotically matching the best bound for gradient descent with well-tuned learning rates. A similar result with convergence rate T^{−1/4} is also shown for stochastic gradient descent.

## Three Mechanisms of Weight Decay Regularization

- Guodong Zhang, Chaoqi Wang, Bowen Xu, Roger Grosse
- abstract@open-review(Poster): Weight decay is one of the standard tricks in the neural network toolbox, but the reasons for its regularization effect are poorly understood, and recent results have cast doubt on the traditional interpretation in terms of $L_2$ regularization. Literal weight decay has been shown to outperform $L_2$ regularization for optimizers for which they differ. We empirically investigate weight decay for three optimization algorithms (SGD, Adam, and K-FAC) and a variety of network architectures. We identify three distinct mechanisms by which weight decay exerts a regularization effect, depending on the particular optimization algorithm and architecture: (1) increasing the effective learning rate, (2) approximately regularizing the input-output Jacobian norm, and (3) reducing the effective damping coefficient for second-order optimization. Our results provide insight into how to improve the regularization of neural networks.

## Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet

- Wieland Brendel, Matthias Bethge
- abstract@open-review(Poster): Deep Neural Networks (DNNs) excel on many complex perceptual tasks but it has proven notoriously difficult to understand how they reach their decisions. We here introduce a high-performance DNN architecture on ImageNet whose decisions are considerably easier to explain. Our model, a simple variant of the ResNet-50 architecture called BagNet, classifies an image based on the occurrences of small local image features without taking into account their spatial ordering. This strategy is closely related to the bag-of-feature (BoF) models popular before the onset of deep learning and reaches a surprisingly high accuracy on ImageNet (87.6% top-5 for 32 x 32 px features and Alexnet performance for 16 x16 px features). The constraint on local features makes it straight-forward to analyse how exactly each part of the image influences the classification. Furthermore, the BagNets behave similar to state-of-the art deep neural networks such as VGG-16, ResNet-152 or DenseNet-169 in terms of feature sensitivity, error distribution and interactions between image parts. This suggests that the improvements of DNNs over previous bag-of-feature classifiers in the last few years is mostly achieved by better fine-tuning rather than by qualitatively different decision strategies.

## Learning Exploration Policies for Navigation

- Tao Chen, Saurabh Gupta, Abhinav Gupta
- abstract@open-review(Poster): Numerous past works have tackled the problem of task-driven navigation. But, how to effectively explore a new environment to enable a variety of down-stream tasks has received much less attention. In this work, we study how agents can autonomously explore realistic and complex 3D environments without the context of task-rewards. We propose a learning-based approach and investigate different policy architectures, reward functions, and training paradigms. We find that use of policies with spatial memory that are bootstrapped with imitation learning and finally finetuned with coverage rewards derived purely from on-board sensors can be effective at exploring novel environments. We show that our learned exploration policies can explore better than classical approaches based on geometry alone and generic learning-based exploration techniques. Finally, we also show how such task-agnostic exploration can be used for down-stream tasks. Videos are available at https://sites.google.com/view/exploration-for-nav/.

## The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks

- Jonathan Frankle, Michael Carbin
- abstract@open-review(Oral): Neural network pruning techniques can reduce the parameter counts of trained networks by over 90%, decreasing storage requirements and improving computational performance of inference without compromising accuracy. However, contemporary experience is that the sparse architectures produced by pruning are difficult to train from the start, which would similarly improve training performance.

We find that a standard pruning technique naturally uncovers subnetworks whose initializations made them capable of training effectively. Based on these results, we articulate the "lottery ticket hypothesis:" dense, randomly-initialized, feed-forward networks contain subnetworks ("winning tickets") that - when trained in isolation - reach test accuracy comparable to the original network in a similar number of iterations. The winning tickets we find have won the initialization lottery: their connections have initial weights that make training particularly effective.

We present an algorithm to identify winning tickets and a series of experiments that support the lottery ticket hypothesis and the importance of these fortuitous initializations. We consistently find winning tickets that are less than 10-20% of the size of several fully-connected and convolutional feed-forward architectures for MNIST and CIFAR10. Above this size, the winning tickets that we find learn faster than the original network and reach higher test accuracy.

## Learning Programmatically Structured Representations with Perceptor Gradients

- Svetlin Penkov, Subramanian Ramamoorthy
- abstract@open-review(Poster): We present the perceptor gradients algorithm -- a novel approach to learning symbolic representations based on the idea of decomposing an agent's policy into i) a perceptor network extracting symbols from raw observation data and ii) a task encoding program which maps the input symbols to output actions. We show that the proposed algorithm is able to learn representations that can be directly fed into a Linear-Quadratic Regulator (LQR) or a general purpose A* planner. Our experimental results confirm that the perceptor gradients algorithm is able to efficiently learn transferable symbolic representations as well as generate new observations according to a semantically meaningful specification.

## Learning Mixed-Curvature Representations in Product Spaces

- Albert Gu, Frederic Sala, Beliz Gunel, Christopher Ré
- abstract@open-review(Poster): The quality of the representations achieved by embeddings is determined by how well the geometry of the embedding space matches the structure of the data. Euclidean space has been the workhorse for embeddings; recently hyperbolic and spherical spaces have gained popularity due to their ability to better embed new types of structured data---such as hierarchical data---but most data is not structured so uniformly. We address this problem by proposing learning embeddings in a product manifold combining multiple copies of these model spaces (spherical, hyperbolic, Euclidean), providing a space of heterogeneous curvature suitable for a wide variety of structures. We introduce a heuristic to estimate the sectional

curvature of graph data and directly determine an appropriate signature---the number of component spaces and their dimensions---of the product manifold. Empirically, we jointly learn the curvature and the embedding in the product space via Riemannian optimization. We discuss how to define and compute intrinsic quantities such as means---a challenging notion for product manifolds---and provably learnable optimization functions. On a range of datasets and reconstruction tasks, our product space embeddings outperform single Euclidean or hyperbolic spaces used in previous works, reducing distortion by 32.55% on a Facebook social network dataset. We learn word embeddings and find that a product of hyperbolic spaces in 50 dimensions consistently improves on baseline Euclidean and hyperbolic embeddings, by 2.6 points in Spearman rank correlation on similarity tasks and 3.4 points on analogy accuracy.

## Stable Recurrent Models

- John Miller, Moritz Hardt
- abstract@open-review(Poster): Stability is a fundamental property of dynamical systems, yet to this date it has had little bearing on the practice of recurrent neural networks. In this work, we conduct a thorough investigation of stable recurrent models. Theoretically, we prove stable recurrent neural networks are well approximated by feed-forward networks for the purpose of both inference and training by gradient descent. Empirically, we demonstrate stable recurrent models often perform as well as their unstable counterparts on benchmark sequence tasks. Taken together, these findings shed light on the effective power of recurrent networks and suggest much of sequence learning happens, or can be made to happen, in the stable regime. Moreover, our results help to explain why in many cases practitioners succeed in replacing recurrent models by feed-forward models.

## Deep Anomaly Detection with Outlier Exposure

- Dan Hendrycks, Mantas Mazeika, Thomas Dietterich
- abstract@open-review(Poster): It is important to detect anomalous inputs when deploying machine learning systems. The use of larger and more complex inputs in deep learning magnifies the difficulty of distinguishing between anomalous and in-distribution examples. At the same time, diverse image and text data are available in enormous quantities. We propose leveraging these data to improve deep anomaly detection by training anomaly detectors against an auxiliary dataset of outliers, an approach we call Outlier Exposure (OE). This enables anomaly detectors to generalize and detect unseen anomalies. In extensive experiments on natural language processing and small- and large-scale vision tasks, we find that Outlier Exposure significantly improves detection performance. We also observe that cutting-edge generative models trained on CIFAR-10 may assign higher likelihoods to SVHN images than to CIFAR-10 images; we use OE to mitigate this issue. We also analyze the flexibility and robustness of Outlier Exposure, and identify characteristics of the auxiliary dataset that improve performance.

## Exemplar Guided Unsupervised Image-to-Image Translation with Semantic Consistency

- Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, Luc Van Gool
- abstract@open-review(Poster): Image-to-image translation has recently received significant attention due to advances in deep learning. Most works focus on learning either a one-to-one mapping in an unsupervised way or a many-to-many mapping in a supervised way. However, a more practical setting is many-to-many mapping in an unsupervised way, which is harder due to the lack of supervision and the complex inner- and cross-domain variations. To alleviate these issues, we propose the Exemplar Guided & Semantically Consistent Image-to-image Translation (EGSC-IT) network which conditions the translation process on an exemplar image in the target domain. We assume that an image comprises of a content component which is shared across domains, and a style component specific to each domain. Under the guidance of an exemplar from the target domain we apply Adaptive Instance Normalization to the shared content component, which allows us to transfer the style information of the target domain to the source domain. To avoid semantic inconsistencies during translation that naturally appear due to the large inner- and cross-domain variations, we introduce the concept of feature masks that provide coarse semantic guidance without requiring the use of any semantic labels. Experimental results on various datasets show that EGSC-IT does not only translate the source image to diverse instances in the target domain, but also preserves the semantic consistency during the process.

## Pay Less Attention with Lightweight and Dynamic Convolutions

- Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, Michael Auli
- abstract@open-review(Oral): Self-attention is a useful mechanism to build generative models for language and images. It determines the importance of context elements by comparing each element to the current time step. In this paper, we show that a very lightweight convolution can perform competitively to the best reported self-attention results. Next, we introduce dynamic convolutions which are simpler and more efficient than self-attention. We predict separate convolution kernels based solely on the current time-step in order to determine the importance of context elements. The number of operations required by this approach scales linearly in the input length, whereas self-attention is quadratic. Experiments on large-scale machine translation, language modeling and abstractive summarization show that dynamic convolutions improve over strong self-attention models. On the WMT'14 English-German test set dynamic convolutions achieve a new state of the art of 29.7 BLEU.

## Doubly Reparameterized Gradient Estimators for Monte Carlo Objectives

- George Tucker, Dieterich Lawson, Shixiang Gu, Chris J. Maddison
- abstract@open-review(Poster): Deep latent variable models have become a popular model choice due to the scalable learning algorithms introduced by (Kingma & Welling 2013, Rezende et al. 2014). These approaches maximize a variational lower bound on the intractable log likelihood of the observed data. Burda et al. (2015) introduced a multi-sample variational bound, IWAE, that is at least as tight as the standard variational lower bound and becomes increasingly tight as the number of samples increases. Counterintuitively, the typical inference network gradient estimator for the IWAE bound performs poorly as the number of samples increases (Rainforth et al. 2018, Le et al. 2018). Roeder et a. (2017) propose an improved gradient estimator, however, are unable to show it is unbiased. We show that it is in fact biased and that the bias can be estimated efficiently with a second application of the reparameterization trick. The doubly reparameterized gradient (DReG) estimator does not suffer as the number of samples increases, resolving the previously raised issues. The same idea can be used to improve many recently introduced training techniques for latent variable models. In particular, we show that this estimator reduces the variance of the IWAE gradient, the reweighted wake-sleep update (RWS) (Bornschein & Bengio 2014), and the jackknife variational inference (JVI) gradient (Nowozin 2018). Finally, we show that this computationally efficient, drop-in estimator translates to improved performance for all three objectives on several modeling tasks.

## Adversarial Attacks on Graph Neural Networks via Meta Learning

- Daniel Zügner, Stephan Günnemann
- abstract@open-review(Poster): Deep learning models for graphs have advanced the state of the art on many tasks. Despite their recent success, little is known about their robustness. We investigate training time attacks on graph neural networks for node classification that perturb the discrete graph structure. Our core principle is to use meta-gradients to solve the bilevel problem underlying training-time attacks, essentially treating the graph as a hyperparameter to optimize. Our experiments show that small graph perturbations consistently lead to a strong decrease in performance for graph convolutional networks, and even transfer to unsupervised embeddings. Remarkably, the perturbations created by our algorithm can misguide the graph neural networks such that they perform worse than a simple baseline that ignores all relational information. Our attacks do not assume any knowledge about or access to the target classifiers.

## Adaptive Gradient Methods with Dynamic Bound of Learning Rate

- Liangchen Luo, Yuanhao Xiong, Yan Liu, Xu Sun
- abstract@open-review(Poster): Adaptive optimization methods such as AdaGrad, RMSprop and Adam have been proposed to achieve a rapid training process with an element-wise scaling term on learning rates. Though prevailing, they are observed to generalize poorly compared with SGD or even fail to converge due to unstable and extreme learning rates. Recent work has put forward some algorithms such as AMSGrad to tackle this issue but they failed to achieve considerable improvement over existing methods. In our paper, we demonstrate that extreme learning rates can lead to poor performance. We provide new variants of Adam and AMSGrad, called AdaBound and AMSBound respectively, which employ dynamic bounds on learning rates to achieve a gradual and smooth transition from adaptive methods to SGD and give a theoretical proof of convergence. We further conduct experiments on various popular tasks and models, which is often insufficient in previous work. Experimental results show that new variants can eliminate the generalization gap between adaptive methods and SGD and maintain higher learning speed early in training at the same time. Moreover, they can bring significant improvement over their prototypes, especially on complex deep networks. The implementation of the algorithm can be found at https://github.com/Luolc/AdaBound .

## Random mesh projectors for inverse problems

- Konik Kothari, *Sidharth Gupta*, Maarten v. de Hoop, Ivan Dokmanic
- abstract@open-review(Poster): We propose a new learning-based approach to solve ill-posed inverse problems in imaging. We address the case where ground truth training samples are rare and the problem is severely ill-posed---both because of the underlying physics and because we can only get few measurements. This setting is common in geophysical imaging and remote sensing. We show that in this case the common approach to directly learn the mapping from the measured data to the reconstruction becomes unstable. Instead, we propose to first learn an ensemble of simpler mappings from the data to projections of the unknown image into random piecewise-constant subspaces. We then combine the projections to form a final reconstruction by solving a deconvolution-like problem. We show experimentally that the proposed method is more robust to measurement noise and corruptions not seen during training than a directly learned inverse.

## A Statistical Approach to Assessing Neural Network Robustness

- Stefan Webb, Tom Rainforth, Yee Whye Teh, M. Pawan Kumar
- abstract@open-review(Poster): We present a new approach to assessing the robustness of neural networks based on estimating the proportion of inputs for which a property is violated. Specifically, we estimate the probability of the event that the property is violated under an input model. Our approach critically varies from the formal verification framework in that when the property can be violated, it provides an informative notion of how robust the network is, rather than just the conventional assertion that the network is not verifiable. Furthermore, it provides an ability to scale to larger networks than formal verification approaches. Though the framework still provides a formal guarantee of satisfiability whenever it successfully finds one or more violations, these advantages do come at the cost of only providing a statistical estimate of unsatisfiability whenever no violation is found. Key to the practical success of our approach is an adaptation of multi-level splitting, a Monte Carlo approach for estimating the probability of rare events, to our statistical robustness framework. We demonstrate that our approach is able to emulate formal verification procedures on benchmark problems, while scaling to larger networks and providing reliable additional information in the form of accurate estimates of the violation probability.

## Learning Actionable Representations with Goal Conditioned Policies

- Dibya Ghosh, Abhishek Gupta, Sergey Levine
- abstract@open-review(Poster): Representation learning is a central challenge across a range of machine learning areas. In reinforcement learning, effective and functional representations have the potential to tremendously accelerate learning progress and solve more challenging problems. Most prior work on representation learning has focused on generative approaches, learning representations that capture all the underlying factors of variation in the observation space in a more disentangled or well-ordered manner. In this paper, we instead aim to learn functionally salient representations: representations that are not necessarily complete in terms of capturing all factors of variation in the observation space, but rather aim to capture those factors of variation that are important for decision making -- that are "actionable". These representations are aware of the dynamics of the environment, and capture only the elements of the observation that are necessary for decision making rather than all factors of variation, eliminating the need for explicit reconstruction. We show how these learned representations can be useful to improve exploration for sparse reward problems, to enable long horizon hierarchical reinforcement learning, and as a state representation for learning policies for downstream tasks. We evaluate our method on a number of simulated environments, and compare it to prior methods for representation learning, exploration, and hierarchical reinforcement learning.

## Learning Implicitly Recurrent CNNs Through Parameter Sharing

- Pedro Savarese, Michael Maire
- abstract@open-review(Poster): We introduce a parameter sharing scheme, in which different layers of a convolutional neural network (CNN) are defined by a learned linear combination of parameter tensors from a global bank of templates. Restricting the number of templates yields a flexible hybridization of traditional CNNs and recurrent networks. Compared to traditional CNNs, we demonstrate substantial parameter savings on standard image classification tasks, while maintaining accuracy. Our simple parameter sharing scheme, though defined via soft weights, in practice often yields trained networks with near strict recurrent structure; with negligible side effects, they convert into networks with actual loops. Training these networks thus implicitly involves discovery of suitable recurrent architectures. Though considering only the aspect of recurrent links, our trained networks achieve accuracy competitive with those built using state-of-the-art neural architecture search (NAS) procedures. Our hybridization of recurrent and convolutional networks may also represent a beneficial architectural bias. Specifically, on synthetic tasks which are algorithmic in nature, our hybrid networks both train faster and extrapolate better to test examples outside the span of the training set.

## Transfer Learning for Sequences via Learning to Collocate

- Wanyun Cui, Guangyu Zheng, Zhiqiang Shen, Sihang Jiang, Wei Wang
- abstract@open-review(Poster): Transfer learning aims to solve the data sparsity for a specific domain by applying information of another domain. Given a sequence (e.g. a natural language sentence), the transfer learning, usually enabled by recurrent neural network (RNN), represent the sequential information transfer. RNN uses a chain of repeating cells to model the sequence data. However, previous studies of neural network based transfer learning simply transfer the information across the whole layers, which are unfeasible for seq2seq and sequence labeling. Meanwhile, such layer-wise transfer learning mechanisms also lose the fine-grained cell-level information from the source domain.

In this paper, we proposed the aligned recurrent transfer, ART, to achieve cell-level information transfer. ART is in a recurrent manner that different cells share the same parameters. Besides transferring the corresponding information at the same position, ART transfers information from all collocated words in the source domain. This strategy enables ART to capture the word collocation across domains in a more flexible way. We conducted extensive experiments on both sequence labeling tasks (POS tagging, NER) and sentence classification (sentiment analysis). ART outperforms the state-of-the-arts over all experiments.

## How to train your MAML

- Antreas Antoniou, Harrison Edwards, Amos Storkey

- abstract@open-review(Poster): The field of few-shot learning has recently seen substantial advancements. Most of these advancements came from casting few-shot learning as a meta-learning problem.Model Agnostic Meta Learning or MAML is currently one of the best approaches for few-shot learning via meta-learning. MAML is simple, elegant and very powerful, however, it has a variety of issues, such as being very sensitive to neural network architectures, often leading to instability during training, requiring arduous hyperparameter searches to stabilize training and achieve high generalization and being very computationally expensive at both training and inference times. In this paper, we propose various modifications to MAML that not only stabilize the system, but also substantially improve the generalization performance, convergence speed and computational overhead of MAML, which we call MAML++.

## Variational Discriminator Bottleneck: Improving Imitation Learning, Inverse RL, and GANs by Constraining Information Flow

- Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel, Sergey Levine
- abstract@open-review(Poster): Adversarial learning methods have been proposed for a wide range of applications, but the training of adversarial models can be notoriously unstable. Effectively balancing the performance of the generator and discriminator is critical, since a discriminator that achieves very high accuracy will produce relatively uninformative gradients. In this work, we propose a simple and general technique to constrain information flow in the discriminator by means of an information bottleneck. By enforcing a constraint on the mutual information between the observations and the discriminator's internal representation, we can effectively modulate the discriminator's accuracy and maintain useful and informative gradients. We demonstrate that our proposed variational discriminator bottleneck (VDB) leads to significant improvements across three distinct application areas for adversarial learning algorithms. Our primary evaluation studies the applicability of the VDB to imitation learning of dynamic continuous control skills, such as running. We show that our method can learn such skills directly from raw video demonstrations, substantially outperforming prior adversarial imitation learning methods. The VDB can also be combined with adversarial inverse reinforcement learning to learn parsimonious reward functions that can be transferred and re-optimized in new settings. Finally, we demonstrate that VDB can train GANs more effectively for image generation, improving upon a number of prior stabilization methods.

## Learning what and where to attend

- Drew Linsley, Dan Shiebler, Sven Eberhardt, Thomas Serre
- abstract@open-review(Poster): Most recent gains in visual recognition have originated from the inclusion of attention mechanisms in deep convolutional networks (DCNs). Because these networks are optimized for object recognition, they learn where to attend using only a weak form of supervision derived from image class labels. Here, we demonstrate the benefit of using stronger supervisory signals by teaching DCNs to attend to image regions that humans deem important for object recognition. We first describe a large-scale online experiment (ClickMe) used to supplement ImageNet with nearly half a million human-derived "top-down" attention maps. Using human psychophysics, we confirm that the identified top-down features from ClickMe are more diagnostic than "bottom-up" saliency features for rapid image categorization. As a proof of concept, we extend a state-of-the-art attention network and demonstrate that adding ClickMe supervision significantly improves its accuracy and yields visual features that are more interpretable and more similar to those used by human observers.

## Learning protein sequence embeddings using information from structure

- Tristan Bepler, Bonnie Berger
- abstract@open-review(Poster): Inferring the structural properties of a protein from its amino acid sequence is a challenging yet important problem in biology. Structures are not known for the vast majority of protein sequences, but structure is critical for understanding function. Existing approaches for detecting structural similarity between proteins from sequence are unable to recognize and exploit structural patterns when sequences have diverged too far, limiting our ability to transfer knowledge between structurally related proteins. We newly approach this problem through the lens of representation learning. We introduce a framework that maps any protein sequence to a sequence of vector embeddings --- one per amino acid position --- that encode structural information. We train bidirectional long short-term memory (LSTM) models on protein sequences with a two-part feedback mechanism that incorporates information from (i) global structural similarity between proteins and (ii) pairwise residue contact maps for individual proteins. To enable learning from structural similarity information, we define a novel similarity measure between arbitrary-length sequences of vector embeddings based on a soft symmetric alignment (SSA) between them. Our method is able to learn useful position-specific embeddings despite lacking direct observations of position-level correspondence between sequences. We show empirically that our multi-task framework outperforms other sequence-based methods and even a top-performing structure-based alignment method when predicting structural similarity, our goal. Finally, we demonstrate that our learned embeddings can be transferred to other protein sequence problems, improving the state-of-the-art in transmembrane domain prediction.

## What do you learn from context? Probing for sentence structure in contextualized word representations

- Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, Ellie Pavlick
- abstract@open-review(Poster): Contextualized representation models such as ELMo (Peters et al., 2018a) and BERT (Devlin et al., 2018) have recently achieved state-of-the-art results on a diverse array of downstream NLP tasks. Building on recent token-level probing work, we introduce a novel edge probing task design and construct a broad suite of sub-sentence tasks derived from the traditional structured NLP pipeline. We probe word-level contextual representations from four recent models and investigate how they encode sentence structure across a range of syntactic, semantic, local, and long-range phenomena. We find that existing models trained on language modeling and translation produce strong representations for syntactic phenomena, but only offer comparably small improvements on semantic tasks over a non-contextual baseline.

## Temporal Difference Variational Auto-Encoder

- Karol Gregor, George Papamakarios, Frederic Besse, Lars Buesing, Theophane Weber
- abstract@open-review(Oral): To act and plan in complex environments, we posit that agents should have a mental simulator of the world with three characteristics: (a) it should build an abstract state representing the condition of the world; (b) it should form a belief which represents uncertainty on the world; (c) it should go beyond simple step-by-step simulation, and exhibit temporal abstraction. Motivated by the absence of a model satisfying all these requirements, we propose TD-VAE, a generative sequence model that learns representations containing explicit beliefs about states several steps into the future, and that can be rolled out directly without single-step transitions. TD-VAE is trained on pairs of temporally separated time points, using an analogue of temporal difference learning used in reinforcement learning.

## Deep learning generalizes because the parameter-function map is biased towards simple functions

- Guillermo Valle-Perez, Chico Q. Camargo, Ard A. Louis
- abstract@open-review(Poster): Deep neural networks (DNNs) generalize remarkably well without explicit regularization even in the strongly over-parametrized regime where classical learning theory would instead predict that they would severely overfit. While many proposals for some kind of implicit regularization have been made to rationalise this success, there is no consensus for the fundamental reason why DNNs do not strongly overfit. In this paper, we provide a new explanation. By applying a very general probability-complexity bound recently derived from algorithmic information theory (AIT), we argue that the parameter-function map of many DNNs should be exponentially biased towards simple functions. We then provide clear evidence for this strong simplicity bias in a model DNN for Boolean functions, as well as in much larger fully connected and convolutional networks trained on CIFAR10 and MNIST. As the target functions in many real problems are expected to be highly structured, this intrinsic simplicity bias helps explain why

deep networks generalize well on real world problems. This picture also facilitates a novel PAC-Bayes approach where the prior is taken over the DNN input-output function space, rather than the more conventional prior over parameter space. If we assume that the training algorithm samples parameters close to uniformly within the zero-error region then the PAC-Bayes theorem can be used to guarantee good expected generalization for target functions producing high-likelihood training sets. By exploiting recently discovered connections between DNNs and Gaussian processes to estimate the marginal likelihood, we produce relatively tight generalization PAC-Bayes error bounds which correlate well with the true error on realistic datasets such as MNIST and CIFAR10 and for architectures including convolutional and fully connected networks.

## CAMOU: Learning Physical Vehicle Camouflages to Adversarially Attack Detectors in the Wild

- Yang Zhang, Hassan Foroosh, Philip David, Boqing Gong
- abstract@open-review(Poster): In this paper, we conduct an intriguing experimental study about the physical adversarial attack on object detectors in the wild. In particular, we learn a camouflage pattern to hide vehicles from being detected by state-of-the-art convolutional neural network based detectors. Our approach alternates between two threads. In the first, we train a neural approximation function to imitate how a simulator applies a camouflage to vehicles and how a vehicle detector performs given images of the camouflaged vehicles. In the second, we minimize the approximated detection score by searching for the optimal camouflage. Experiments show that the learned camouflage can not only hide a vehicle from the image-based detectors under many test cases but also generalizes to different environments, vehicles, and object detectors.

## MAE: Mutual Posterior-Divergence Regularization for Variational AutoEncoders

- Xuezhe Ma, Chunting Zhou, Eduard Hovy
- abstract@open-review(Poster): Variational Autoencoder (VAE), a simple and effective deep generative model, has led to a number of impressive empirical successes and spawned many advanced variants and theoretical investigations. However, recent studies demonstrate that, when equipped with expressive generative distributions (aka. decoders), VAE suffers from learning uninformative latent representations with the observation called KL Varnishing, in which case VAE collapses into an unconditional generative model. In this work, we introduce mutual posterior-divergence regularization, a novel regularization that is able to control the geometry of the latent space to accomplish meaningful representation learning, while achieving comparable or superior capability of density estimation.Experiments on three image benchmark datasets demonstrate that, when equipped with powerful decoders, our model performs well both on density estimation and representation learning.

## A Convergence Analysis of Gradient Descent for Deep Linear Neural Networks

- Sanjeev Arora, Nadav Cohen, Noah Golowich, Wei Hu
- abstract@open-review(Poster): We analyze speed of convergence to global optimum for gradient descent training a deep linear neural network by minimizing the L2 loss over whitened data. Convergence at a linear rate is guaranteed when the following hold: (i) dimensions of hidden layers are at least the minimum of the input and output dimensions; (ii) weight matrices at initialization are approximately balanced; and (iii) the initial loss is smaller than the loss of any rank-deficient solution. The assumptions on initialization (conditions (ii) and (iii)) are necessary, in the sense that violating any one of them may lead to convergence failure. Moreover, in the important case of output dimension 1, i.e. scalar regression, they are met, and thus convergence to global optimum holds, with constant probability under a random initialization scheme. Our results significantly extend previous analyses, e.g., of deep linear residual networks (Bartlett et al., 2018).

## Don't Settle for Average, Go for the Max: Fuzzy Sets and Max-Pooled Word Vectors

- Vitalii Zhelezniak, Aleksandar Savkov, April Shen, Francesco Moramarco, Jack Flann, Nils Y. Hammerla
- abstract@open-review(Poster): Recent literature suggests that averaged word vectors followed by simple post-processing outperform many deep learning methods on semantic textual similarity tasks. Furthermore, when averaged word vectors are trained supervised on large corpora of paraphrases, they achieve state-of-the-art results on standard STS benchmarks. Inspired by these insights, we push the limits of word embeddings even further. We propose a novel fuzzy bag-of-words (FBoW) representation for text that contains all the words in the vocabulary simultaneously but with different degrees of membership, which are derived from similarities between word vectors. We show that max-pooled word vectors are only a special case of fuzzy BoW and should be compared via fuzzy Jaccard index rather than cosine similarity. Finally, we propose DynaMax, a completely unsupervised and non-parametric similarity measure that dynamically extracts and max-pools good features depending on the sentence pair. This method is both efficient and easy to implement, yet outperforms current baselines on STS tasks by a large margin and is even competitive with supervised word vectors trained to directly optimise cosine similarity.

## The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision

- Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, Jiajun Wu
- abstract@open-review(Oral): We propose the Neuro-Symbolic Concept Learner (NS-CL), a model that learns visual concepts, words, and semantic parsing of sentences without explicit supervision on any of them; instead, our model learns by simply looking at images and reading paired questions and answers. Our model builds an object-based scene representation and translates sentences into executable, symbolic programs. To bridge the learning of two modules, we use a neuro-symbolic reasoning module that executes these programs on the latent scene representation. Analogical to human concept learning, the perception module learns visual concepts based on the language description of the object being referred to. Meanwhile, the learned visual concepts facilitate learning new words and parsing new sentences. We use curriculum learning to guide the searching over the large compositional space of images and language. Extensive experiments demonstrate the accuracy and efficiency of our model on learning visual concepts, word representations, and semantic parsing of sentences. Further, our method allows easy generalization to new object attributes, compositions, language concepts, scenes and questions, and even new program domains. It also empowers applications including visual question answering and bidirectional image-text retrieval.

## The role of over-parametrization in generalization of neural networks

- Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, Nathan Srebro
- abstract@open-review(Poster): Despite existing work on ensuring generalization of neural networks in terms of scale sensitive complexity measures, such as norms, margin and sharpness, these complexity measures do not offer an explanation of why neural networks generalize better with over-parametrization. In this work we suggest a novel complexity measure based on unit-wise capacities resulting in a tighter generalization bound for two layer ReLU networks. Our capacity bound correlates with the behavior of test error with increasing network sizes (within the range reported in the experiments), and could partly explain the improvement in generalization with over-parametrization. We further present a matching lower bound for the Rademacher complexity that improves over previous capacity lower bounds for neural networks.

## On the Convergence of A Class of Adam-Type Algorithms for Non-Convex Optimization

- Xiangyi Chen, Sijia Liu, Ruoyu Sun, Mingyi Hong
- abstract@open-review(Poster): This paper studies a class of adaptive gradient based momentum algorithms that update the search directions and learning rates simultaneously using past gradients. This class, which we refer to as the "``Adam-type", includes the popular algorithms such as Adam, AMSGrad, AdaGrad. Despite their popularity in training deep neural networks (DNNs), the convergence of these algorithms for solving non-convex problems

remains an open question. In this paper, we develop an analysis framework and a set of mild sufficient conditions that guarantee the convergence of the Adam-type methods, with a convergence rate of order $O(\log{T}/\sqrt{T})$ for non-convex stochastic optimization. Our convergence analysis applies to a new algorithm called AdaFom (AdaGrad with First Order Momentum). We show that the conditions are essential, by identifying concrete examples in which violating the conditions makes an algorithm diverge. Besides providing one of the first comprehensive analysis for Adam-type methods in the non-convex setting, our results can also help the practitioners to easily monitor the progress of algorithms and determine their convergence behavior.

## Gradient descent aligns the layers of deep linear networks

- Ziwei Ji, Matus Telgarsky
- abstract@open-review(Poster): This paper establishes risk convergence and asymptotic weight matrix alignment --- a form of implicit regularization --- of gradient flow and gradient descent when applied to deep linear networks on linearly separable data. In more detail, for gradient flow applied to strictly decreasing loss functions (with similar results for gradient descent with particular decreasing step sizes): (i) the risk converges to 0; (ii) the normalized i-th weight matrix asymptotically equals its rank-1 approximation u_iv_i^T; (iii) these rank-1 matrices are aligned across layers, meaning |v_{i+1}^T u_i| -> 1. In the case of the logistic loss (binary cross entropy), more can be said: the linear function induced by the network --- the product of its weight matrices --- converges to the same direction as the maximum margin solution. This last property was identified in prior work, but only under assumptions on gradient descent which here are implied by the alignment phenomenon.

## Hierarchical RL Using an Ensemble of Proprioceptive Periodic Policies

- Kenneth Marino, Abhinav Gupta, Rob Fergus, Arthur Szlam
- abstract@open-review(Poster): In this paper we introduce a simple, robust approach to hierarchically training an agent in the setting of sparse reward tasks. The agent is split into a low-level and a high-level policy. The low-level policy only accesses internal, proprioceptive dimensions of the state observation. The low-level policies are trained with a simple reward that encourages changing the values of the non-proprioceptive dimensions. Furthermore, it is induced to be periodic with the use a ``phase function." The high-level policy is trained using a sparse, task-dependent reward, and operates by choosing which of the low-level policies to run at any given time. Using this approach, we solve difficult maze and navigation tasks with sparse rewards using the Mujoco Ant and Humanoid agents and show improvement over recent hierarchical methods.

## Learning To Simulate

- Nataniel Ruiz, Samuel Schulter, Manmohan Chandraker
- abstract@open-review(Poster): Simulation is a useful tool in situations where training data for machine learning models is costly to annotate or even hard to acquire. In this work, we propose a reinforcement learning-based method for automatically adjusting the parameters of any (non-differentiable) simulator, thereby controlling the distribution of synthesized data in order to maximize the accuracy of a model trained on that data. In contrast to prior art that hand-crafts these simulation parameters or adjusts only parts of the available parameters, our approach fully controls the simulator with the actual underlying goal of maximizing accuracy, rather than mimicking the real data distribution or randomly generating a large volume of data. We find that our approach (i) quickly converges to the optimal simulation parameters in controlled experiments and (ii) can indeed discover good sets of parameters for an image rendering simulator in actual computer vision applications.

## Multilingual Neural Machine Translation With Soft Decoupled Encoding

- Xinyi Wang, Hieu Pham, Philip Arthur, Graham Neubig
- abstract@open-review(Poster): Multilingual training of neural machine translation (NMT) systems has led to impressive accuracy improvements on low-resource languages. However, there are still significant challenges in efficiently learning word representations in the face of paucity of data. In this paper, we propose Soft Decoupled Encoding (SDE), a multilingual lexicon encoding framework specifically designed to share lexical-level information intelligently without requiring heuristic preprocessing such as pre-segmenting the data. SDE represents a word by its spelling through a character encoding, and its semantic meaning through a latent embedding space shared by all languages. Experiments on a standard dataset of four low-resource languages show consistent improvements over strong multilingual NMT baselines, with gains of up to 2 BLEU on one of the tested languages, achieving the new state-of-the-art on all four language pairs.

## Meta-Learning with Latent Embedding Optimization

- Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, Raia Hadsell
- abstract@open-review(Poster): Gradient-based meta-learning techniques are both widely applicable and proficient at solving challenging few-shot learning and fast adaptation problems. However, they have practical difficulties when operating on high-dimensional parameter spaces in extreme low-data regimes. We show that it is possible to bypass these limitations by learning a data-dependent latent generative representation of model parameters, and performing gradient-based meta-learning in this low-dimensional latent space. The resulting approach, latent embedding optimization (LEO), decouples the gradient-based adaptation procedure from the underlying high-dimensional space of model parameters. Our evaluation shows that LEO can achieve state-of-the-art performance on the competitive miniImageNet and tieredImageNet few-shot classification tasks. Further analysis indicates LEO is able to capture uncertainty in the data, and can perform adaptation more effectively by optimizing in latent space.

## Visual Semantic Navigation using Scene Priors

- Wei Yang, Xiaolong Wang, Ali Farhadi, Abhinav Gupta, Roozbeh Mottaghi
- abstract@open-review(Poster): How do humans navigate to target objects in novel scenes? Do we use the semantic/functional priors we have built over years to efficiently search and navigate? For example, to search for mugs, we search cabinets near the coffee machine and for fruits we try the fridge. In this work, we focus on incorporating semantic priors in the task of semantic navigation. We propose to use Graph Convolutional Networks for incorporating the prior knowledge into a deep reinforcement learning framework. The agent uses the features from the knowledge graph to predict the actions. For evaluation, we use the AI2-THOR framework. Our experiments show how semantic knowledge improves the performance significantly. More importantly, we show improvement in generalization to unseen scenes and/or objects.

## AdaShift: Decorrelation and Convergence of Adaptive Learning Rate Methods

- Zhiming Zhou, *Qingru Zhang*, Guansong Lu, Hongwei Wang, Weinan Zhang, Yong Yu
- abstract@open-review(Poster): Adam is shown not being able to converge to the optimal solution in certain cases. Researchers recently propose several algorithms to avoid the issue of non-convergence of Adam, but their efficiency turns out to be unsatisfactory in practice. In this paper, we provide a new insight into the non-convergence issue of Adam as well as other adaptive learning rate methods. We argue that there exists an inappropriate correlation between gradient $g_t$ and the second moment term $v_t$ in Adam ($t$ is the timestep), which results in that a large gradient is likely to have small step size while a small gradient may have a large step size. We demonstrate that such unbalanced step sizes are the fundamental cause of non-convergence of Adam, and we further prove that decorrelating $v_t$ and $g_t$ will lead to unbiased step size for each gradient, thus solving the non-convergence problem of Adam. Finally, we propose AdaShift, a novel adaptive learning rate method that decorrelates $v_t$ and $g_t$ by temporal shifting, i.e., using

temporally shifted gradient $g_{t-n}$ to calculate $v_t$. The experiment results demonstrate that AdaShift is able to address the non-convergence issue of Adam, while still maintaining a competitive performance with Adam in terms of both training speed and generalization.

## Sliced Wasserstein Auto-Encoders

- Soheil Kolouri, Phillip E. Pope, Charles E. Martin, Gustavo K. Rohde
- abstract@open-review(Poster): In this paper we use the geometric properties of the optimal transport (OT) problem and the Wasserstein distances to define a prior distribution for the latent space of an auto-encoder. We introduce Sliced-Wasserstein Auto-Encoders (SWAE), that enable one to shape the distribution of the latent space into any samplable probability distribution without the need for training an adversarial network or having a likelihood function specified. In short, we regularize the auto-encoder loss with the sliced-Wasserstein distance between the distribution of the encoded training samples and a samplable prior distribution. We show that the proposed formulation has an efficient numerical solution that provides similar capabilities to Wasserstein Auto-Encoders (WAE) and Variational Auto-Encoders (VAE), while benefiting from an embarrassingly simple implementation. We provide extensive error analysis for our algorithm, and show its merits on three benchmark datasets.

## Amortized Bayesian Meta-Learning

- Sachin Ravi, Alex Beatson
- abstract@open-review(Poster): Meta-learning, or learning-to-learn, has proven to be a successful strategy in attacking problems in supervised learning and reinforcement learning that involve small amounts of data. State-of-the-art solutions involve learning an initialization and/or learning algorithm using a set of training episodes so that the meta learner can generalize to an evaluation episode quickly. These methods perform well but often lack good quantification of uncertainty, which can be vital to real-world applications when data is lacking. We propose a meta-learning method which efficiently amortizes hierarchical variational inference across tasks, learning a prior distribution over neural network weights so that a few steps of Bayes by Backprop will produce a good task-specific approximate posterior. We show that our method produces good uncertainty estimates on contextual bandit and few-shot learning benchmarks.

## Local SGD Converges Fast and Communicates Little

- Sebastian U. Stich
- abstract@open-review(Poster): Mini-batch stochastic gradient descent (SGD) is state of the art in large scale distributed training. The scheme can reach a linear speed-up with respect to the number of workers, but this is rarely seen in practice as the scheme often suffers from large network delays and bandwidth limits. To overcome this communication bottleneck recent works propose to reduce the communication frequency. An algorithm of this type is local SGD that runs SGD independently in parallel on different workers and averages the sequences only once in a while. This scheme shows promising results in practice, but eluded thorough theoretical analysis.

We prove concise convergence rates for local SGD on convex problems and show that it converges at the same rate as mini-batch SGD in terms of number of evaluated gradients, that is, the scheme achieves linear speed-up in the number of workers and mini-batch size. The number of communication rounds can be reduced up to a factor of T^{1/2}---where T denotes the number of total steps---compared to mini-batch SGD. This also holds for asynchronous implementations.

Local SGD can also be used for large scale training of deep learning models. The results shown here aim serving as a guideline to further explore the theoretical and practical aspects of local SGD in these applications.

## Learning Protein Structure with a Differentiable Simulator

- John Ingraham, Adam Riesselman, Chris Sander, Debora Marks
- abstract@open-review(Oral): The Boltzmann distribution is a natural model for many systems, from brains to materials and biomolecules, but is often of limited utility for fitting data because Monte Carlo algorithms are unable to simulate it in available time. This gap between the expressive capabilities and sampling practicalities of energy-based models is exemplified by the protein folding problem, since energy landscapes underlie contemporary knowledge of protein biophysics but computer simulations are challenged to fold all but the smallest proteins from first principles. In this work we aim to bridge the gap between the expressive capacity of energy functions and the practical capabilities of their simulators by using an unrolled Monte Carlo simulation as a model for data. We compose a neural energy function with a novel and efficient simulator based on Langevin dynamics to build an end-to-end-differentiable model of atomic protein structure given amino acid sequence information. We introduce techniques for stabilizing backpropagation under long roll-outs and demonstrate the model's capacity to make multimodal predictions and to, in some cases, generalize to unobserved protein fold types when trained on a large corpus of protein structures.

## textTOvec: DEEP CONTEXTUALIZED NEURAL AUTOREGRESSIVE TOPIC MODELS OF LANGUAGE WITH DISTRIBUTED COMPOSITIONAL PRIOR

- Pankaj Gupta, Yatin Chaudhary, Florian Buettner, Hinrich Schuetze
- abstract@open-review(Poster): We address two challenges of probabilistic topic modelling in order to better estimate the probability of a word in a given context, i.e., P(wordjcontext) : (1) No Language Structure in Context: Probabilistic topic models ignore word order by summarizing a given context as a “bag-of-word” and consequently the semantics of words in the context is lost. In this work, we incorporate language structure by combining a neural autoregressive topic model (TM) with a LSTM based language model (LSTM-LM) in a single probabilistic framework. The LSTM-LM learns a vector-space representation of each word by accounting for word order in local collocation patterns, while the TM simultaneously learns a latent representation from the entire document. In addition, the LSTM-LM models complex characteristics of language (e.g., syntax and semantics), while the TM discovers the underlying thematic structure in a collection of documents. We unite two complementary paradigms of learning the meaning of word occurrences by combining a topic model and a language model in a unified probabilistic framework, named as ctx-DocNADE. (2) Limited Context and/or Smaller training corpus of documents: In settings with a small number of word occurrences (i.e., lack of context) in short text or data sparsity in a corpus of few documents, the application of TMs is challenging. We address this challenge by incorporating external knowledge into neural autoregressive topic models via a language modelling approach: we use word embeddings as input of a LSTM-LM with the aim to improve the wordtopic mapping on a smaller and/or short-text corpus. The proposed DocNADE extension is named as ctx-DocNADEe.

We present novel neural autoregressive topic model variants coupled with neural language models and embeddings priors that consistently outperform state-of-theart generative topic models in terms of generalization (perplexity), interpretability (topic coherence) and applicability (retrieval and classification) over 6 long-text and 8 short-text datasets from diverse domains.

## Neural Program Repair by Jointly Learning to Localize and Repair

- Marko Vasic, Aditya Kanade, Petros Maniatis, David Bieber, Rishabh Singh
- abstract@open-review(Poster): Due to its potential to improve programmer productivity and software quality, automated program repair has been an active topic of research. Newer techniques harness neural networks to learn directly from examples of buggy programs and their fixes. In this work, we consider a recently identified class of bugs called variable-misuse bugs. The state-of-the-art solution for variable misuse enumerates potential fixes for all

possible bug locations in a program, before selecting the best prediction. We show that it is beneficial to train a model that jointly and directly localizes and repairs variable-misuse bugs. We present multi-headed pointer networks for this purpose, with one head each for localization and repair. The experimental results show that the joint model significantly outperforms an enumerative solution that uses a pointer based model for repair alone.

## Human-level Protein Localization with Convolutional Neural Networks

- Elisabeth Rumetshofer, Markus Hofmarcher, Clemens Röhrl, Sepp Hochreiter, Günter Klambauer
- abstract@open-review(Poster): Localizing a specific protein in a human cell is essential for understanding cellular functions and biological processes of underlying diseases. A promising, low-cost,and time-efficient biotechnology for localizing proteins is high-throughput fluorescence microscopy imaging (HTI). This imaging technique stains the protein of interest in a cell with fluorescent antibodies and subsequently takes a microscopic image. Together with images of other stained proteins or cell organelles and the annotation by the Human Protein Atlas project, these images provide a rich source of information on the protein location which can be utilized by computational methods. It is yet unclear how precise such methods are and whether they can compete with human experts. We here focus on deep learning image analysis methods and, in particular, on Convolutional Neural Networks (CNNs)since they showed overwhelming success across different imaging tasks. We pro-pose a novel CNN architecture “GapNet-PL” that has been designed to tackle the characteristics of HTI data and uses global averages of filters at different abstraction levels. We present the largest comparison of CNN architectures including GapNet-PL for protein localization in HTI images of human cells. GapNet-PL outperforms all other competing methods and reaches close to perfect localization in all 13 tasks with an average AUC of 98% and F1 score of 78%. On a separate test set the performance of GapNet-PL was compared with three human experts and 25 scholars. GapNet-PL achieved an accuracy of 91%, significantly (p-value 1.1e−6) outperforming the best human expert with an accuracy of 72%.

## From Language to Goals: Inverse Reinforcement Learning for Vision-Based Instruction Following

- Justin Fu, Anoop Korattikara, Sergey Levine, Sergio Guadarrama
- abstract@open-review(Poster): Reinforcement learning is a promising framework for solving control problems, but its use in practical situations is hampered by the fact that reward functions are often difficult to engineer. Specifying goals and tasks for autonomous machines, such as robots, is a significant challenge: conventionally, reward functions and goal states have been used to communicate objectives. But people can communicate objectives to each other simply by describing or demonstrating them. How can we build learning algorithms that will allow us to tell machines what we want them to do? In this work, we investigate the problem of grounding language commands as reward functions using inverse reinforcement learning, and argue that language-conditioned rewards are more transferable than language-conditioned policies to new environments. We propose language-conditioned reward learning (LC-RL), which grounds language commands as a reward function represented by a deep neural network. We demonstrate that our model learns rewards that transfer to novel tasks and environments on realistic, high-dimensional visual environments with natural language commands, whereas directly learning a language-conditioned policy leads to poor performance.

## ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech

- Wei Ping, Kainan Peng, Jitong Chen
- abstract@open-review(Poster): In this work, we propose a new solution for parallel wave generation by WaveNet. In contrast to parallel WaveNet (van Oord et al., 2018), we distill a Gaussian inverse autoregressive flow from the autoregressive WaveNet by minimizing a regularized KL divergence between their highly-peaked output distributions. Our method computes the KL divergence in closed-form, which simplifies the training algorithm and provides very efficient distillation. In addition, we introduce the first text-to-wave neural architecture for speech synthesis, which is fully convolutional and enables fast end-to-end training from scratch. It significantly outperforms the previous pipeline that connects a text-to-spectrogram model to a separately trained WaveNet (Ping et al., 2018). We also successfully distill a parallel waveform synthesizer conditioned on the hidden representation in this end-to-end model.

## Variational Autoencoder with Arbitrary Conditioning

- Oleg Ivanov, Michael Figurnov, Dmitry Vetrov
- abstract@open-review(Poster): We propose a single neural probabilistic model based on variational autoencoder that can be conditioned on an arbitrary subset of observed features and then sample the remaining features in "one shot". The features may be both real-valued and categorical. Training of the model is performed by stochastic variational Bayes. The experimental evaluation on synthetic data, as well as feature imputation and image inpainting problems, shows the effectiveness of the proposed approach and diversity of the generated samples.

## Janossy Pooling: Learning Deep Permutation-Invariant Functions for Variable-Size Inputs

- Ryan L. Murphy, Balasubramaniam Srinivasan, Vinayak Rao, Bruno Ribeiro
- abstract@open-review(Poster): We consider a simple and overarching representation for permutation-invariant functions of sequences (or set functions). Our approach, which we call Janossy pooling, expresses a permutation-invariant function as the average of a permutation-sensitive function applied to all reorderings of the input sequence. This allows us to leverage the rich and mature literature on permutation-sensitive functions to construct novel and flexible permutation-invariant functions. If carried out naively, Janossy pooling can be computationally prohibitive. To allow computational tractability, we consider three kinds of approximations: canonical orderings of sequences, functions with k-order interactions, and stochastic optimization algorithms with random permutations. Our framework unifies a variety of existing work in the literature, and suggests possible modeling and algorithmic extensions. We explore a few in our experiments, which demonstrate improved performance over current state-of-the-art methods.

## The Deep Weight Prior

- Andrei Atanov, Arsenii Ashukha, Kirill Struminsky, Dmitriy Vetrov, Max Welling
- abstract@open-review(Poster): Bayesian inference is known to provide a general framework for incorporating prior knowledge or specific properties into machine learning models via carefully choosing a prior distribution. In this work, we propose a new type of prior distributions for convolutional neural networks, deep weight prior (DWP), that exploit generative models to encourage a specific structure of trained convolutional filters e.g., spatial correlations of weights. We define DWP in the form of an implicit distribution and propose a method for variational inference with such type of implicit priors. In experiments, we show that DWP improves the performance of Bayesian neural networks when training data are limited, and initialization of weights with samples from DWP accelerates training of conventional convolutional neural networks.

## A new dog learns old tricks: RL finds classic optimization algorithms

- Weiwei Kong, Christopher Liaw, Aranyak Mehta, D. Sivakumar
- abstract@open-review(Poster): This paper introduces a novel framework for learning algorithms to solve online combinatorial optimization problems. Towards this goal, we introduce a number of key ideas from traditional algorithms and complexity theory. First, we draw a new connection between primal-dual methods and reinforcement learning. Next, we introduce the concept of adversarial distributions (universal and high-entropy training sets), which are distributions that encourage the learner to find algorithms that work well in the worst case. We test our new ideas on a number of optimization

problem such as the AdWords problem, the online knapsack problem, and the secretary problem. Our results indicate that the models have learned behaviours that are consistent with the traditional optimal algorithms for these problems.

## DOM-Q-NET: Grounded RL on Structured Language

- Sheng Jia, Jamie Ryan Kiros, Jimmy Ba
- abstract@open-review(Poster): Building agents to interact with the web would allow for significant improvements in knowledge understanding and representation learning. However, web navigation tasks are difficult for current deep reinforcement learning (RL) models due to the large discrete action space and the varying number of actions between the states. In this work, we introduce DOM-Q-NET, a novel architecture for RL-based web navigation to address both of these problems. It parametrizes Q functions with separate networks for different action categories: clicking a DOM element and typing a string input. Our model utilizes a graph neural network to represent the tree-structured HTML of a standard web page. We demonstrate the capabilities of our model on the MiniWoB environment where we can match or outperform existing work without the use of expert demonstrations. Furthermore, we show 2x improvements in sample efficiency when training in the multi-task setting, allowing our model to transfer learned behaviours across tasks.

## InstaGAN: Instance-aware Image-to-Image Translation

- Sangwoo Mo, Minsu Cho, Jinwoo Shin
- abstract@open-review(Poster): Unsupervised image-to-image translation has gained considerable attention due to the recent impressive progress based on generative adversarial networks (GANs). However, previous methods often fail in challenging cases, in particular, when an image has multiple target instances and a translation task involves significant changes in shape, e.g., translating pants to skirts in fashion images. To tackle the issues, we propose a novel method, coined instance-aware GAN (InstaGAN), that incorporates the instance information (e.g., object segmentation masks) and improves multi-instance transfiguration. The proposed method translates both an image and the corresponding set of instance attributes while maintaining the permutation invariance property of the instances. To this end, we introduce a context preserving loss that encourages the network to learn the identity function outside of target instances. We also propose a sequential mini-batch inference/training technique that handles multiple instances with a limited GPU memory and enhances the network to generalize better for multiple instances. Our comparative evaluation demonstrates the effectiveness of the proposed method on different image datasets, in particular, in the aforementioned challenging cases. Code and results are available in https://github.com/sangwoomo/instagan

## Learning to Describe Scenes with Programs

- Yunchao Liu, Zheng Wu, Daniel Ritchie, William T. Freeman, Joshua B. Tenenbaum, Jiajun Wu
- abstract@open-review(Poster): Human scene perception goes beyond recognizing a collection of objects and their pairwise relations. We understand higher-level, abstract regularities within the scene such as symmetry and repetition. Current vision recognition modules and scene representations fall short in this dimension. In this paper, we present scene programs, representing a scene via a symbolic program for its objects, attributes, and their relations. We also propose a model that infers such scene programs by exploiting a hierarchical, object-based scene representation. Experiments demonstrate that our model works well on synthetic data and transfers to real images with such compositional structure. The use of scene programs has enabled a number of applications, such as complex visual analogy-making and scene extrapolation.

## Data-Dependent Coresets for Compressing Neural Networks with Applications to Generalization Bounds

- Cenk Baykal, Lucas Liebenwein, Igor Gilitschenski, Dan Feldman, Daniela Rus
- abstract@open-review(Poster): We present an efficient coresets-based neural network compression algorithm that sparsifies the parameters of a trained fully-connected neural network in a manner that provably approximates the network's output. Our approach is based on an importance sampling scheme that judiciously defines a sampling distribution over the neural network parameters, and as a result, retains parameters of high importance while discarding redundant ones. We leverage a novel, empirical notion of sensitivity and extend traditional coreset constructions to the application of compressing parameters. Our theoretical analysis establishes guarantees on the size and accuracy of the resulting compressed network and gives rise to generalization bounds that may provide new insights into the generalization properties of neural networks. We demonstrate the practical effectiveness of our algorithm on a variety of neural network configurations and real-world data sets.

## InfoBot: Transfer and Exploration via the Information Bottleneck

- Anirudh Goyal, Riashat Islam, DJ Strouse, Zafarali Ahmed, Hugo Larochelle, Matthew Botvinick, Yoshua Bengio, Sergey Levine
- abstract@open-review(Poster): A central challenge in reinforcement learning is discovering effective policies for tasks where rewards are sparsely distributed. We postulate that in the absence of useful reward signals, an effective exploration strategy should seek out {\it decision states}. These states lie at critical junctions in the state space from where the agent can transition to new, potentially unexplored regions. We propose to learn about decision states from prior experience. By training a goal-conditioned model with an information bottleneck, we can identify decision states by examining where the model accesses the goal state through the bottleneck. We find that this simple mechanism effectively identifies decision states, even in partially observed settings. In effect, the model learns the sensory cues that correlate with potential subgoals. In new environments, this model can then identify novel subgoals for further exploration, guiding the agent through a sequence of potential decision states and through new regions of the state space.

## Emerging Disentanglement in Auto-Encoder Based Unsupervised Image Content Transfer

- Ori Press, Tomer Galanti, Sagie Benaim, Lior Wolf
- abstract@open-review(Poster): We study the problem of learning to map, in an unsupervised way, between domains $A$ and $B$, such that the samples $\vb \in B$ contain all the information that exists in samples $\va\in A$ and some additional information. For example, ignoring occlusions, $B$ can be people with glasses, $A$ people without, and the glasses, would be the added information. When mapping a sample $\va$ from the first domain to the other domain, the missing information is replicated from an independent reference sample $\vb\in B$. Thus, in the above example, we can create, for every person without glasses a version with the glasses observed in any face image.

Our solution employs a single two-pathway encoder and a single decoder for both domains. The common part of the two domains and the separate part are encoded as two vectors, and the separate part is fixed at zero for domain $A$. The loss terms are minimal and involve reconstruction losses for the two domains and a domain confusion term. Our analysis shows that under mild assumptions, this architecture, which is much simpler than the literature guided-translation methods, is enough to ensure disentanglement between the two domains. We present convincing results in a few visual domains, such as no-glasses to glasses, adding facial hair based on a reference image, etc.

## Learning Embeddings into Entropic Wasserstein Spaces

- Charlie Frogner, Farzaneh Mirzazadeh, Justin Solomon
- abstract@open-review(Poster): Despite their prevalence, Euclidean embeddings of data are fundamentally limited in their ability to capture latent semantic structures, which need not conform to Euclidean spatial assumptions. Here we consider an alternative, which embeds data as discrete probability distributions in a Wasserstein space, endowed with an optimal transport metric. Wasserstein spaces are much larger and more flexible than Euclidean spaces, in that they can successfully embed a wider variety of metric structures. We propose to exploit this flexibility by learning an embedding that captures the semantic information in the Wasserstein distance between embedded distributions. We examine empirically the representational capacity of

such learned Wasserstein embeddings, showing that they can embed a wide variety of complex metric structures with smaller distortion than an equivalent Euclidean embedding. We also investigate an application to word embedding, demonstrating a unique advantage of Wasserstein embeddings: we can directly visualize the high-dimensional embedding, as it is a probability distribution on a low-dimensional space. This obviates the need for dimensionality reduction techniques such as t-SNE for visualization.

## Wasserstein Barycenter Model Ensembling

- Pierre Dognin, *Igor Melnyk*, Youssef Mroueh, *Jarret Ross*, Cicero Dos Santos, *Tom Sercu*
- abstract@open-review(Poster): In this paper we propose to perform model ensembling in a multiclass or a multilabel learning setting using Wasserstein (W.) barycenters. Optimal transport metrics, such as the Wasserstein distance, allow incorporating semantic side information such as word embeddings. Using W. barycenters to find the consensus between models allows us to balance confidence and semantics in finding the agreement between the models. We show applications of Wasserstein ensembling in attribute-based classification, multilabel learning and image captioning generation. These results show that the W. ensembling is a viable alternative to the basic geometric or arithmetic mean ensembling.

## Generalizable Adversarial Training via Spectral Normalization

- Farzan Farnia, Jesse Zhang, David Tse
- abstract@open-review(Poster): Deep neural networks (DNNs) have set benchmarks on a wide array of supervised learning tasks. Trained DNNs, however, often lack robustness to minor adversarial perturbations to the input, which undermines their true practicality. Recent works have increased the robustness of DNNs by fitting networks using adversarially-perturbed training samples, but the improved performance can still be far below the performance seen in non-adversarial settings. A significant portion of this gap can be attributed to the decrease in generalization performance due to adversarial training. In this work, we extend the notion of margin loss to adversarial settings and bound the generalization error for DNNs trained under several well-known gradient-based attack schemes, motivating an effective regularization scheme based on spectral normalization of the DNN's weight matrices. We also provide a computationally-efficient method for normalizing the spectral norm of convolutional layers with arbitrary stride and padding schemes in deep convolutional networks. We evaluate the power of spectral normalization extensively on combinations of datasets, network architectures, and adversarial training schemes.

## Unsupervised Speech Recognition via Segmental Empirical Output Distribution Matching

- Chih-Kuan Yeh, Jianshu Chen, Chengzhu Yu, Dong Yu
- abstract@open-review(Poster): We consider the problem of training speech recognition systems without using any labeled data, under the assumption that the learner can only access to the input utterances and a phoneme language model estimated from a non-overlapping corpus. We propose a fully unsupervised learning algorithm that alternates between solving two sub-problems: (i) learn a phoneme classifier for a given set of phoneme segmentation boundaries, and (ii) refining the phoneme boundaries based on a given classifier. To solve the first sub-problem, we introduce a novel unsupervised cost function named Segmental Empirical Output Distribution Matching, which generalizes the work in (Liu et al., 2017) to segmental structures. For the second sub-problem, we develop an approximate MAP approach to refining the boundaries obtained from Wang et al. (2017). Experimental results on TIMIT dataset demonstrate the success of this fully unsupervised phoneme recognition system, which achieves a phone error rate (PER) of 41.6%. Although it is still far away from the state-of-the-art supervised systems, we show that with oracle boundaries and matching language model, the PER could be improved to 32.5%. This performance approaches the supervised system of the same model architecture, demonstrating the great potential of the proposed method.

## Learning when to Communicate at Scale in Multiagent Cooperative and Competitive Tasks

- Amanpreet Singh, Tushar Jain, Sainbayar Sukhbaatar
- abstract@open-review(Poster): Learning when to communicate and doing that effectively is essential in multi-agent tasks. Recent works show that continuous communication allows efficient training with back-propagation in multi-agent scenarios, but have been restricted to fully-cooperative tasks. In this paper, we present Individualized Controlled Continuous Communication Model (IC3Net) which has better training efficiency than simple continuous communication model, and can be applied to semi-cooperative and competitive settings along with the cooperative settings. IC3Net controls continuous communication with a gating mechanism and uses individualized rewards foreach agent to gain better performance and scalability while fixing credit assignment issues. Using variety of tasks including StarCraft BroodWars explore and combat scenarios, we show that our network yields improved performance and convergence rates than the baselines as the scale increases. Our results convey that IC3Net agents learn when to communicate based on the scenario and profitability.

## ProxQuant: Quantized Neural Networks via Proximal Operators

- Yu Bai, Yu-Xiang Wang, Edo Liberty
- abstract@open-review(Poster): To make deep neural networks feasible in resource-constrained environments (such as mobile devices), it is beneficial to quantize models by using low-precision weights. One common technique for quantizing neural networks is the straight-through gradient method, which enables back-propagation through the quantization mapping. Despite its empirical success, little is understood about why the straight-through gradient method works. Building upon a novel observation that the straight-through gradient method is in fact identical to the well-known Nesterov's dual-averaging algorithm on a quantization constrained optimization problem, we propose a more principled alternative approach, called ProxQuant , that formulates quantized network training as a regularized learning problem instead and optimizes it via the prox-gradient method. ProxQuant does back-propagation on the underlying full-precision vector and applies an efficient prox-operator in between stochastic gradient steps to encourage quantizedness. For quantizing ResNets and LSTMs, ProxQuant outperforms state-of-the-art results on binary quantization and is on par with state-of-the-art on multi-bit quantization. We further perform theoretical analyses showing that ProxQuant converges to stationary points under mild smoothness assumptions, whereas variants such as lazy prox-gradient method can fail to converge in the same setting.

## Optimal Completion Distillation for Sequence Learning

- Sara Sabour, William Chan, Mohammad Norouzi
- abstract@open-review(Poster): We present Optimal Completion Distillation (OCD), a training procedure for optimizing sequence to sequence models based on edit distance. OCD is efficient, has no hyper-parameters of its own, and does not require pre-training or joint optimization with conditional log-likelihood. Given a partial sequence generated by the model, we first identify the set of optimal suffixes that minimize the total edit distance, using an efficient dynamic programming algorithm. Then, for each position of the generated sequence, we use a target distribution which puts equal probability on the first token of all the optimal suffixes. OCD achieves the state-of-the-art performance on end-to-end speech recognition, on both Wall Street Journal and Librispeech datasets, achieving $9.3\%$ WER and $4.5\%$ WER, respectively.

## Feature Intertwiner for Object Detection

- Hongyang Li, Bo Dai, Shaoshuai Shi, Wanli Ouyang, Xiaogang Wang

- abstract@open-review(Poster): A well-trained model should classify objects with unanimous score for every category. This requires the high-level semantic features should be alike among samples, despite a wide span in resolution, texture, deformation, etc. Previous works focus on re-designing the loss function or proposing new regularization constraints on the loss. In this paper, we address this problem via a new perspective. For each category, it is assumed that there are two sets in the feature space: one with more reliable information and the other with less reliable source. We argue that the reliable set could guide the feature learning of the less reliable set during training - in spirit of student mimicking teacher's behavior and thus pushing towards a more compact class centroid in the high-dimensional space. Such a scheme also benefits the reliable set since samples become more closer within the same category - implying that it is easilier for the classifier to identify. We refer to this mutual learning process as feature intertwiner and embed the spirit into object detection. It is well-known that objects of low resolution are more difficult to detect due to the loss of detailed information during network forward pass. We thus regard objects of high resolution as the reliable set and objects of low resolution as the less reliable set. Specifically, an intertwiner is achieved by minimizing the distribution divergence between two sets. We design a historical buffer to represent all previous samples in the reliable set and utilize them to guide the feature learning of the less reliable set. The design of obtaining an effective feature representation for the reliable set is further investigated, where we introduce the optimal transport (OT) algorithm into the framework. Samples in the less reliable set are better aligned with the reliable set with aid of OT metric. Incorporated with such a plug-and-play intertwiner, we achieve an evident improvement over previous state-of-the-arts on the COCO object detection benchmark.

## Diversity and Depth in Per-Example Routing Models

- Prajit Ramachandran, Quoc V. Le
- abstract@open-review(Poster): Routing models, a form of conditional computation where examples are routed through a subset of components in a larger network, have shown promising results in recent works. Surprisingly, routing models to date have lacked important properties, such as architectural diversity and large numbers of routing decisions. Both architectural diversity and routing depth can increase the representational power of a routing network. In this work, we address both of these deficiencies. We discuss the significance of architectural diversity in routing models, and explain the tradeoffs between capacity and optimization when increasing routing depth. In our experiments, we find that adding architectural diversity to routing models significantly improves performance, cutting the error rates of a strong baseline by 35% on an Omniglot setup. However, when scaling up routing depth, we find that modern routing techniques struggle with optimization. We conclude by discussing both the positive and negative results, and suggest directions for future research.

## Learning concise representations for regression by evolving networks of trees

- William La Cava, Tilak Raj Singh, James Taggart, Srinivas Suri, Jason H. Moore
- abstract@open-review(Poster): We propose and study a method for learning interpretable representations for the task of regression. Features are represented as networks of multi-type expression trees comprised of activation functions common in neural networks in addition to other elementary functions. Differentiable features are trained via gradient descent, and the performance of features in a linear model is used to weight the rate of change among subcomponents of each representation. The search process maintains an archive of representations with accuracy-complexity trade-offs to assist in generalization and interpretation. We compare several stochastic optimization approaches within this framework. We benchmark these variants on 100 open-source regression problems in comparison to state-of-the-art machine learning approaches. Our main finding is that this approach produces the highest average test scores across problems while producing representations that are orders of magnitude smaller than the next best performing method (gradient boosting). We also report a negative result in which attempts to directly optimize the disentanglement of the representation result in more highly correlated features.

## FFJORD: Free-Form Continuous Dynamics for Scalable Reversible Generative Models

- Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, David Duvenaud
- abstract@open-review(Oral): A promising class of generative models maps points from a simple distribution to a complex distribution through an invertible neural network. Likelihood-based training of these models requires restricting their architectures to allow cheap computation of Jacobian determinants. Alternatively, the Jacobian trace can be used if the transformation is specified by an ordinary differential equation. In this paper, we use Hutchinson's trace estimator to give a scalable unbiased estimate of the log-density. The result is a continuous-time invertible generative model with unbiased density estimation and one-pass sampling, while allowing unrestricted neural network architectures. We demonstrate our approach on high-dimensional density estimation, image generation, and variational inference, achieving the state-of-the-art among exact likelihood methods with efficient sampling.

## Exploration by random network distillation

- Yuri Burda, Harrison Edwards, Amos Storkey, Oleg Klimov
- abstract@open-review(Poster): We introduce an exploration bonus for deep reinforcement learning methods that is easy to implement and adds minimal overhead to the computation performed. The bonus is the error of a neural network predicting features of the observations given by a fixed randomly initialized neural network. We also introduce a method to flexibly combine intrinsic and extrinsic rewards. We find that the random network distillation (RND) bonus combined with this increased flexibility enables significant progress on several hard exploration Atari games. In particular we establish state of the art performance on Montezuma's Revenge, a game famously difficult for deep reinforcement learning methods. To the best of our knowledge, this is the first method that achieves better than average human performance on this game without using demonstrations or having access the underlying state of the game, and occasionally completes the first level. This suggests that relatively simple methods that scale well can be sufficient to tackle challenging exploration problems.

## Hierarchical Generative Modeling for Controllable Speech Synthesis

- Wei-Ning Hsu, Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen, Patrick Nguyen, Ruoming Pang
- abstract@open-review(Poster): This paper proposes a neural end-to-end text-to-speech (TTS) model which can control latent attributes in the generated speech that are rarely annotated in the training data, such as speaking style, accent, background noise, and recording conditions. The model is formulated as a conditional generative model with two levels of hierarchical latent variables. The first level is a categorical variable, which represents attribute groups (e.g. clean/noisy) and provides interpretability. The second level, conditioned on the first, is a multivariate Gaussian variable, which characterizes specific attribute configurations (e.g. noise level, speaking rate) and enables disentangled fine-grained control over these attributes. This amounts to using a Gaussian mixture model (GMM) for the latent distribution. Extensive evaluation demonstrates its ability to control the aforementioned attributes. In particular, it is capable of consistently synthesizing high-quality clean speech regardless of the quality of the training data for the target speaker.

## Generative Question Answering: Learning to Answer the Whole Question

- Mike Lewis, Angela Fan
- abstract@open-review(Poster): Discriminative question answering models can overfit to superficial biases in datasets, because their loss function saturates when any clue makes the answer likely. We introduce generative models of the joint distribution of questions and answers, which are trained to explain the whole question, not just to answer it.Our question answering (QA) model is implemented by learning a prior over answers, and a conditional language model to generate the question given the answer—allowing scalable and interpretable many-hop reasoning as the question is generated word-by-word.

Our model achieves competitive performance with specialised discriminative models on the SQUAD and CLEVR benchmarks, indicating that it is a more general architecture for language understanding and reasoning than previous work. The model greatly improves generalisation both from biased training data and to adversarial testing data, achieving a new state-of-the-art on ADVERSARIAL SQUAD. We will release our code.

## Decoupled Weight Decay Regularization

- Ilya Loshchilov, Frank Hutter
- abstract@open-review(Poster): L$_2$ regularization and weight decay regularization are equivalent for standard stochastic gradient descent (when rescaled by the learning rate), but as we demonstrate this is \emph{not} the case for adaptive gradient algorithms, such as Adam. While common implementations of these algorithms employ L$_2$ regularization (often calling it ``weight decay'' in what may be misleading due to the inequivalence we expose), we propose a simple modification to recover the original formulation of weight decay regularization by \emph{decoupling} the weight decay from the optimization steps taken w.r.t. the loss function. We provide empirical evidence that our proposed modification (i) decouples the optimal choice of weight decay factor from the setting of the learning rate for both standard SGD and Adam and (ii) substantially improves Adam's generalization performance, allowing it to compete with SGD with momentum on image classification datasets (on which it was previously typically outperformed by the latter). Our proposed decoupled weight decay has already been adopted by many researchers, and the community has implemented it in TensorFlow and PyTorch; the complete source code for our experiments is available at \url{https://github.com/loshchil/AdamW-and-SGDW}

## Probabilistic Recursive Reasoning for Multi-Agent Reinforcement Learning

- Ying Wen, Yaodong Yang, Rui Luo, Jun Wang, Wei Pan
- abstract@open-review(Poster): Humans are capable of attributing latent mental contents such as beliefs, or intentions to others. The social skill is critical in everyday life to reason about the potential consequences of their behaviors so as to plan ahead. It is known that humans use this reasoning ability recursively, i.e. considering what others believe about their own beliefs. In this paper, we start from level-$1$ recursion and introduce a probabilistic recursive reasoning (PR2) framework for multi-agent reinforcement learning. Our hypothesis is that it is beneficial for each agent to account for how the opponents would react to its future behaviors. Under the PR2 framework, we adopt variational Bayes methods to approximate the opponents' conditional policy, to which each agent finds the best response and then improve their own policy. We develop decentralized-training-decentralized-execution algorithms, PR2-Q and PR2-Actor-Critic, that are proved to converge in the self-play scenario when there is one Nash equilibrium. Our methods are tested on both the matrix game and the differential game, which have a non-trivial equilibrium where common gradient-based methods fail to converge. Our experiments show that it is critical to reason about how the opponents believe about what the agent believes. We expect our work to contribute a new idea of modeling the opponents to the multi-agent reinforcement learning community.

## Learning to Represent Edits

- Pengcheng Yin, Graham Neubig, Miltiadis Allamanis, Marc Brockschmidt, Alexander L. Gaunt
- abstract@open-review(Poster): We introduce the problem of learning distributed representations of edits. By combining a "neural editor" with an "edit encoder", our models learn to represent the salient information of an edit and can be used to apply edits to new inputs. We experiment on natural language and source code edit data. Our evaluation yields promising results that suggest that our neural network models learn to capture the structure and semantics of edits. We hope that this interesting task and data source will inspire other researchers to work further on this problem.

## Unsupervised Domain Adaptation for Distance Metric Learning

- Kihyuk Sohn, Wenling Shang, Xiang Yu, Manmohan Chandraker
- abstract@open-review(Poster): Unsupervised domain adaptation is a promising avenue to enhance the performance of deep neural networks on a target domain, using labels only from a source domain. However, the two predominant methods, domain discrepancy reduction learning and semi-supervised learning, are not readily applicable when source and target domains do not share a common label space. This paper addresses the above scenario by learning a representation space that retains discriminative power on both the (labeled) source and (unlabeled) target domains while keeping representations for the two domains well-separated. Inspired by a theoretical analysis, we first reformulate the disjoint classification task, where the source and target domains correspond to non-overlapping class labels, to a verification one. To handle both within and cross domain verifications, we propose a Feature Transfer Network (FTN) to separate the target feature space from the original source space while aligned with a transformed source space. Moreover, we present a non-parametric multi-class entropy minimization loss to further boost the discriminative power of FTNs on the target domain. In experiments, we first illustrate how FTN works in a controlled setting of adapting from MNIST-M to MNIST with disjoint digit classes between the two domains and then demonstrate the effectiveness of FTNs through state-of-the-art performances on a cross-ethnicity face recognition problem.

## Bayesian Deep Convolutional Networks with Many Channels are Gaussian Processes

- Roman Novak, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Jiri Hron, Daniel A. Abolafia, Jeffrey Pennington, Jascha Sohl-dickstein
- abstract@open-review(Poster): There is a previously identified equivalence between wide fully connected neural networks (FCNs) and Gaussian processes (GPs). This equivalence enables, for instance, test set predictions that would have resulted from a fully Bayesian, infinitely wide trained FCN to be computed without ever instantiating the FCN, but by instead evaluating the corresponding GP. In this work, we derive an analogous equivalence for multi-layer convolutional neural networks (CNNs) both with and without pooling layers, and achieve state of the art results on CIFAR10 for GPs without trainable kernels. We also introduce a Monte Carlo method to estimate the GP corresponding to a given neural network architecture, even in cases where the analytic form has too many terms to be computationally feasible.

Surprisingly, in the absence of pooling layers, the GPs corresponding to CNNs with and without weight sharing are identical. As a consequence, translation equivariance, beneficial in finite channel CNNs trained with stochastic gradient descent (SGD), is guaranteed to play no role in the Bayesian treatment of the infinite channel limit - a qualitative difference between the two regimes that is not present in the FCN case. We confirm experimentally, that while in some scenarios the performance of SGD-trained finite CNNs approaches that of the corresponding GPs as the channel count increases, with careful tuning SGD-trained CNNs can significantly outperform their corresponding GPs, suggesting advantages from SGD training compared to fully Bayesian parameter estimation.

## Efficient Training on Very Large Corpora via Gramian Estimation

- Walid Krichene, Nicolas Mayoraz, Steffen Rendle, Li Zhang, Xinyang Yi, Lichan Hong, Ed Chi, John Anderson
- abstract@open-review(Poster): We study the problem of learning similarity functions over very large corpora using neural network embedding models. These models are typically trained using SGD with random sampling of unobserved pairs, with a sample size that grows quadratically with the corpus size, making it expensive to scale. We propose new efficient methods to train these models without having to sample unobserved pairs. Inspired by matrix factorization, our approach relies on adding a global quadratic penalty and expressing this term as the inner-product of two generalized Gramians. We show that the gradient of this term can be efficiently computed by maintaining estimates of the Gramians, and develop variance reduction schemes to improve the quality of the estimates. We conduct large-scale experiments that show a significant improvement both in training time and generalization performance compared to sampling methods.

## A comprehensive, application-oriented study of catastrophic forgetting in DNNs

- B. Pfülb, A. Gepperth
- abstract@open-review(Poster): We present a large-scale empirical study of catastrophic forgetting (CF) in modern Deep Neural Network (DNN) models that perform sequential (or: incremental) learning. A new experimental protocol is proposed that takes into account typical constraints encountered in application scenarios. As the investigation is empirical, we evaluate CF behavior on the hitherto largest number of visual classification datasets, from each of which we construct a representative number of Sequential Learning Tasks (SLTs) in close alignment to previous works on CF. Our results clearly indicate that there is no model that avoids CF for all investigated datasets and SLTs under application conditions. We conclude with a discussion of potential solutions and workarounds to CF, notably for the EWC and IMM models.

## Variational Autoencoders with Jointly Optimized Latent Dependency Structure

- Jiawei He, Yu Gong, Joseph Marino, Greg Mori, Andreas Lehrmann
- abstract@open-review(Poster): We propose a method for learning the dependency structure between latent variables in deep latent variable models. Our general modeling and inference framework combines the complementary strengths of deep generative models and probabilistic graphical models. In particular, we express the latent variable space of a variational autoencoder (VAE) in terms of a Bayesian network with a learned, flexible dependency structure. The network parameters, variational parameters as well as the latent topology are optimized simultaneously with a single objective. Inference is formulated via a sampling procedure that produces expectations over latent variable structures and incorporates top-down and bottom-up reasoning over latent variable values. We validate our framework in extensive experiments on MNIST, Omniglot, and CIFAR-10. Comparisons to state-of-the-art structured variational autoencoder baselines show improvements in terms of the expressiveness of the learned model.

## Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset

- Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, Douglas Eck
- abstract@open-review(Oral): Generating musical audio directly with neural networks is notoriously difficult because it requires coherently modeling structure at many different timescales. Fortunately, most music is also highly structured and can be represented as discrete note events played on musical instruments. Herein, we show that by using notes as an intermediate representation, we can train a suite of models capable of transcribing, composing, and synthesizing audio waveforms with coherent musical structure on timescales spanning six orders of magnitude (~0.1 ms to ~100 s), a process we call Wave2Midi2Wave. This large advance in the state of the art is enabled by our release of the new MAESTRO (MIDI and Audio Edited for Synchronous TRacks and Organization) dataset, composed of over 172 hours of virtuosic piano performances captured with fine alignment (~3 ms) between note labels and audio waveforms. The networks and the dataset together present a promising approach toward creating new expressive and interpretable neural models of music.

## Verification of Non-Linear Specifications for Neural Networks

- Chongli Qin, Krishnamurthy (Dj) Dvijotham, Brendan O'Donoghue, Rudy Bunel, Robert Stanforth, Sven Gowal, Jonathan Uesato, Grzegorz Swirszcz, Pushmeet Kohli
- abstract@open-review(Poster): Prior work on neural network verification has focused on specifications that are linear functions of the output of the network, e.g., invariance of the classifier output under adversarial perturbations of the input. In this paper, we extend verification algorithms to be able to certify richer properties of neural networks. To do this we introduce the class of convex-relaxable specifications, which constitute nonlinear specifications that can be verified using a convex relaxation. We show that a number of important properties of interest can be modeled within this class, including conservation of energy in a learned dynamics model of a physical system; semantic consistency of a classifier's output labels under adversarial perturbations and bounding errors in a system that predicts the summation of handwritten digits. Our experimental evaluation shows that our method is able to effectively verify these specifications. Moreover, our evaluation exposes the failure modes in models which cannot be verified to satisfy these specifications. Thus, emphasizing the importance of training models not just to fit training data but also to be consistent with specifications.

## Improving Sequence-to-Sequence Learning via Optimal Transport

- Liqun Chen, Yizhe Zhang, Ruiyi Zhang, Chenyang Tao, Zhe Gan, Haichao Zhang, Bai Li, Dinghan Shen, Changyou Chen, Lawrence Carin
- abstract@open-review(Poster): Sequence-to-sequence models are commonly trained via maximum likelihood estimation (MLE). However, standard MLE training considers a word-level objective, predicting the next word given the previous ground-truth partial sentence. This procedure focuses on modeling local syntactic patterns, and may fail to capture long-range semantic structure. We present a novel solution to alleviate these issues. Our approach imposes global sequence-level guidance via new supervision based on optimal transport, enabling the overall characterization and preservation of semantic features. We further show that this method can be understood as a Wasserstein gradient flow trying to match our model to the ground truth sequence distribution. Extensive experiments are conducted to validate the utility of the proposed approach, showing consistent improvements over a wide variety of NLP tasks, including machine translation, abstractive text summarization, and image captioning.

## LEARNING TO PROPAGATE LABELS: TRANSDUCTIVE PROPAGATION NETWORK FOR FEW-SHOT LEARNING

- Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, Yi Yang
- abstract@open-review(Poster): The goal of few-shot learning is to learn a classifier that generalizes well even when trained with a limited number of training instances per class. The recently introduced meta-learning approaches tackle this problem by learning a generic classifier across a large number of multiclass classification tasks and generalizing the model to a new task. Yet, even with such meta-learning, the low-data problem in the novel classification task still remains. In this paper, we propose Transductive Propagation Network (TPN), a novel meta-learning framework for transductive inference that classifies the entire test set at once to alleviate the low-data problem. Specifically, we propose to learn to propagate labels from labeled instances to unlabeled test instances, by learning a graph construction module that exploits the manifold structure in the data. TPN jointly learns both the parameters of feature embedding and the graph construction in an end-to-end manner. We validate TPN on multiple benchmark datasets, on which it largely outperforms existing few-shot learning approaches and achieves the state-of-the-art results.

## Universal Transformers

- Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, Lukasz Kaiser
- abstract@open-review(Poster): Recurrent neural networks (RNNs) sequentially process data by updating their state with each new data point, and have long been the de facto choice for sequence modeling tasks. However, their inherently sequential computation makes them slow to train. Feed-forward and convolutional architectures have recently been shown to achieve superior results on some sequence modeling tasks such as machine translation, with the added advantage that they concurrently process all inputs in the sequence, leading to easy parallelization and faster training times. Despite these successes, however, popular feed-forward sequence models like the Transformer fail to generalize in many simple tasks that recurrent models handle with ease, e.g. copying strings or even simple logical inference when the string or formula lengths exceed those observed at training time. We propose the Universal Transformer (UT), a parallel-in-time self-attentive recurrent sequence model which can be cast as a generalization of the Transformer model and which addresses these issues. UTs combine the parallelizability and global receptive field of feed-forward sequence models like the Transformer with the recurrent inductive bias of RNNs. We also add a dynamic per-position halting mechanism and find that it improves accuracy on several tasks. In contrast to the standard Transformer, under certain assumptions UTs can be shown to be Turing-complete. Our experiments show that UTs outperform standard Transformers on a wide range of algorithmic and language understanding tasks, including the challenging LAMBADA language modeling task where UTs achieve a new state of the art, and machine translation where UTs achieve a 0.9 BLEU improvement over Transformers on the WMT14 En-De dataset.

## An Empirical study of Binary Neural Networks' Optimisation

- Milad Alizadeh, Javier Fernández-Marqués, Nicholas D. Lane, Yarin Gal
- abstract@open-review(Poster): Binary neural networks using the Straight-Through-Estimator (STE) have been shown to achieve state-of-the-art results, but their training process is not well-founded. This is due to the discrepancy between the evaluated function in the forward path, and the weight updates in the back-propagation, updates which do not correspond to gradients of the forward path. Efficient convergence and accuracy of binary models often rely on careful fine-tuning and various ad-hoc techniques. In this work, we empirically identify and study the effectiveness of the various ad-hoc techniques commonly used in the literature, providing best-practices for efficient training of binary models. We show that adapting learning rates using second moment methods is crucial for the successful use of the STE, and that other optimisers can easily get stuck in local minima. We also find that many of the commonly employed tricks are only effective towards the end of the training, with these methods making early stages of the training considerably slower. Our analysis disambiguates necessary from unnecessary ad-hoc techniques for training of binary neural networks, paving the way for future development of solid theoretical foundations for these. Our newly-found insights further lead to new procedures which make training of existing binary neural networks notably faster.

## A2BCD: Asynchronous Acceleration with Optimal Complexity

- Robert Hannah, Fei Feng, Wotao Yin
- abstract@open-review(Poster): In this paper, we propose the Asynchronous Accelerated Nonuniform Randomized Block Coordinate Descent algorithm (A2BCD). We prove A2BCD converges linearly to a solution of the convex minimization problem at the same rate as NU_ACDM, so long as the maximum delay is not too large. This is the first asynchronous Nesterov-accelerated algorithm that attains any provable speedup. Moreover, we then prove that these algorithms both have optimal complexity. Asynchronous algorithms complete much faster iterations, and A2BCD has optimal complexity. Hence we observe in experiments that A2BCD is the top-performing coordinate descent algorithm, converging up to 4-5x faster than NU_ACDM on some data sets in terms of wall-clock time. To motivate our theory and proof techniques, we also derive and analyze a continuous-time analog of our algorithm and prove it converges at the same rate.

## Backpropamine: training self-modifying neural networks with differentiable neuromodulated plasticity

- Thomas Miconi, Aditya Rawal, Jeff Clune, Kenneth O. Stanley
- abstract@open-review(Poster): The impressive lifelong learning in animal brains is primarily enabled by plastic changes in synaptic connectivity. Importantly, these changes are not passive, but are actively controlled by neuromodulation, which is itself under the control of the brain. The resulting self-modifying abilities of the brain play an important role in learning and adaptation, and are a major basis for biological reinforcement learning. Here we show for the first time that artificial neural networks with such neuromodulated plasticity can be trained with gradient descent. Extending previous work on differentiable Hebbian plasticity, we propose a differentiable formulation for the neuromodulation of plasticity. We show that neuromodulated plasticity improves the performance of neural networks on both reinforcement learning and supervised learning tasks. In one task, neuromodulated plastic LSTMs with millions of parameters outperform standard LSTMs on a benchmark language modeling task (controlling for the number of parameters). We conclude that differentiable neuromodulation of plasticity offers a powerful new framework for training neural networks.

## Discovery of Natural Language Concepts in Individual Units of CNNs

- Seil Na, Yo Joong Choe, Dong-Hyun Lee, Gunhee Kim
- abstract@open-review(Poster): Although deep convolutional networks have achieved improved performance in many natural language tasks, they have been treated as black boxes because they are difficult to interpret. Especially, little is known about how they represent language in their intermediate layers. In an attempt to understand the representations of deep convolutional networks trained on language tasks, we show that individual units are selectively responsive to specific morphemes, words, and phrases, rather than responding to arbitrary and uninterpretable patterns. In order to quantitatively analyze such intriguing phenomenon, we propose a concept alignment method based on how units respond to replicated text. We conduct analyses with different architectures on multiple datasets for classification and translation tasks and provide new insights into how deep models understand natural language.

## Learning Multi-Level Hierarchies with Hindsight

- Andrew Levy, George Konidaris, Robert Platt, Kate Saenko
- abstract@open-review(Poster): Hierarchical agents have the potential to solve sequential decision making tasks with greater sample efficiency than their non-hierarchical counterparts because hierarchical agents can break down tasks into sets of subtasks that only require short sequences of decisions. In order to realize this potential of faster learning, hierarchical agents need to be able to learn their multiple levels of policies in parallel so these simpler subproblems can be solved simultaneously. Yet, learning multiple levels of policies in parallel is hard because it is inherently unstable: changes in a policy at one level of the hierarchy may cause changes in the transition and reward functions at higher levels in the hierarchy, making it difficult to jointly learn multiple levels of policies. In this paper, we introduce a new Hierarchical Reinforcement Learning (HRL) framework, Hierarchical Actor-Critic (HAC), that can overcome the instability issues that arise when agents try to jointly learn multiple levels of policies. The main idea behind HAC is to train each level of the hierarchy independently of the lower levels by training each level as if the lower level policies are already optimal. We demonstrate experimentally in both grid world and simulated robotics domains that our approach can significantly accelerate learning relative to other non-hierarchical and hierarchical methods. Indeed, our framework is the first to successfully learn 3-level hierarchies in parallel in tasks with continuous state and action spaces.

## ProMP: Proximal Meta-Policy Search

- Jonas Rothfuss, Dennis Lee, Ignasi Clavera, Tamim Asfour, Pieter Abbeel
- abstract@open-review(Poster): Credit assignment in Meta-reinforcement learning (Meta-RL) is still poorly understood. Existing methods either neglect credit assignment to pre-adaptation behavior or implement it naively. This leads to poor sample-efficiency during meta-training as well as ineffective task identification strategies. This paper provides a theoretical analysis of credit assignment in gradient-based Meta-RL. Building on the gained insights we develop a novel meta-learning algorithm that overcomes both the issue of poor credit assignment and previous difficulties in estimating meta-policy gradients. By controlling the statistical distance of both pre-adaptation and adapted policies during meta-policy search, the proposed algorithm endows efficient and stable meta-learning. Our approach leads to superior pre-adaptation policy behavior and consistently outperforms previous Meta-RL algorithms in sample-efficiency, wall-clock time, and asymptotic performance.

## Complement Objective Training

- Hao-Yun Chen, Pei-Hsin Wang, Chun-Hao Liu, Shih-Chieh Chang, Jia-Yu Pan, Yu-Ting Chen, Wei Wei, Da-Cheng Juan
- abstract@open-review(Poster): Learning with a primary objective, such as softmax cross entropy for classification and sequence generation, has been the norm for training deep neural networks for years. Although being a widely-adopted approach, using cross entropy as the primary objective exploits mostly the information from the ground-truth class for maximizing data likelihood, and largely ignores information from the complement (incorrect) classes. We argue that, in addition to the primary objective, training also using a complement objective that leverages information from the complement classes can be

effective in improving model performance. This motivates us to study a new training paradigm that maximizes the likelihood of the ground-truth class while neutralizing the probabilities of the complement classes. We conduct extensive experiments on multiple tasks ranging from computer vision to natural language understanding. The experimental results confirm that, compared to the conventional training with just one primary objective, training also with the complement objective further improves the performance of the state-of-the-art models across all tasks. In addition to the accuracy improvement, we also show that models trained with both primary and complement objectives are more robust to single-step adversarial attacks.

## BabyAI: A Platform to Study the Sample Efficiency of Grounded Language Learning

- Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, Yoshua Bengio
- abstract@open-review(Poster): Allowing humans to interactively train artificial agents to understand language instructions is desirable for both practical and scientific reasons. Though, given the lack of sample efficiency in current learning methods, reaching this goal may require substantial research efforts. We introduce the BabyAI research platform, with the goal of supporting investigations towards including humans in the loop for grounded language learning. The BabyAI platform comprises an extensible suite of 19 levels of increasing difficulty. Each level gradually leads the agent towards acquiring a combinatorially rich synthetic language, which is a proper subset of English. The platform also provides a hand-crafted bot agent, which simulates a human teacher. We report estimated amount of supervision required for training neural reinforcement and behavioral-cloning agents on some BabyAI levels. We put forward strong evidence that current deep learning methods are not yet sufficiently sample-efficient in the context of learning a language with compositional properties.

## Slimmable Neural Networks

- Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, Thomas Huang
- abstract@open-review(Poster): We present a simple and general method to train a single neural network executable at different widths (number of channels in a layer), permitting instant and adaptive accuracy-efficiency trade-offs at runtime. Instead of training individual networks with different width configurations, we train a shared network with switchable batch normalization. At runtime, the network can adjust its width on the fly according to on-device benchmarks and resource constraints, rather than downloading and offloading different models. Our trained networks, named slimmable neural networks, achieve similar (and in many cases better) ImageNet classification accuracy than individually trained models of MobileNet v1, MobileNet v2, ShuffleNet and ResNet-50 at different widths respectively. We also demonstrate better performance of slimmable models compared with individual ones across a wide range of applications including COCO bounding-box object detection, instance segmentation and person keypoint detection without tuning hyper-parameters. Lastly we visualize and discuss the learned features of slimmable networks. Code and models are available at: https://github.com/JiahuiYu/slimmable_networks

## Learning Self-Imitating Diverse Policies

- Tanmay Gangwani, Qiang Liu, Jian Peng
- abstract@open-review(Poster): The success of popular algorithms for deep reinforcement learning, such as policy-gradients and Q-learning, relies heavily on the availability of an informative reward signal at each timestep of the sequential decision-making process. When rewards are only sparsely available during an episode, or a rewarding feedback is provided only after episode termination, these algorithms perform sub-optimally due to the difficultly in credit assignment. Alternatively, trajectory-based policy optimization methods, such as cross-entropy method and evolution strategies, do not require per-timestep rewards, but have been found to suffer from high sample complexity by completing forgoing the temporal nature of the problem. Improving the efficiency of RL algorithms in real-world problems with sparse or episodic rewards is therefore a pressing need. In this work, we introduce a self-imitation learning algorithm that exploits and explores well in the sparse and episodic reward settings. We view each policy as a state-action visitation distribution and formulate policy optimization as a divergence minimization problem. We show that with Jensen-Shannon divergence, this divergence minimization problem can be reduced into a policy-gradient algorithm with shaped rewards learned from experience replays. Experimental results indicate that our algorithm works comparable to existing algorithms in environments with dense rewards, and significantly better in environments with sparse and episodic rewards. We then discuss limitations of self-imitation learning, and propose to solve them by using Stein variational policy gradient descent with the Jensen-Shannon kernel to learn multiple diverse policies. We demonstrate its effectiveness on a challenging variant of continuous-control MuJoCo locomotion tasks.

## Graph HyperNetworks for Neural Architecture Search

- Chris Zhang, Mengye Ren, Raquel Urtasun
- abstract@open-review(Poster): Neural architecture search (NAS) automatically finds the best task-specific neural network topology, outperforming many manual architecture designs. However, it can be prohibitively expensive as the search requires training thousands of different networks, while each training run can last for hours. In this work, we propose the Graph HyperNetwork (GHN) to amortize the search cost: given an architecture, it directly generates the weights by running inference on a graph neural network. GHNs model the topology of an architecture and therefore can predict network performance more accurately than regular hypernetworks and premature early stopping. To perform NAS, we randomly sample architectures and use the validation accuracy of networks with GHN generated weights as the surrogate search signal. GHNs are fast - they can search nearly 10× faster than other random search methods on CIFAR-10 and ImageNet. GHNs can be further extended to the anytime prediction setting, where they have found networks with better speed-accuracy tradeoff than the state-of-the-art manual designs.

## Deep Layers as Stochastic Solvers

- Adel Bibi, Bernard Ghanem, Vladlen Koltun, Rene Ranftl
- abstract@open-review(Poster): We provide a novel perspective on the forward pass through a block of layers in a deep network. In particular, we show that a forward pass through a standard dropout layer followed by a linear layer and a non-linear activation is equivalent to optimizing a convex objective with a single iteration of a $\tau$-nice Proximal Stochastic Gradient method. We further show that replacing standard Bernoulli dropout with additive dropout is equivalent to optimizing the same convex objective with a variance-reduced proximal method. By expressing both fully-connected and convolutional layers as special cases of a high-order tensor product, we unify the underlying convex optimization problem in the tensor setting and derive a formula for the Lipschitz constant $L$ used to determine the optimal step size of the above proximal methods. We conduct experiments with standard convolutional networks applied to the CIFAR-10 and CIFAR-100 datasets and show that replacing a block of layers with multiple iterations of the corresponding solver, with step size set via $L$, consistently improves classification accuracy.

## ARM: Augment-REINFORCE-Merge Gradient for Stochastic Binary Networks

- Mingzhang Yin, Mingyuan Zhou
- abstract@open-review(Poster): To backpropagate the gradients through stochastic binary layers, we propose the augment-REINFORCE-merge (ARM) estimator that is unbiased, exhibits low variance, and has low computational complexity. Exploiting variable augmentation, REINFORCE, and reparameterization, the ARM estimator achieves adaptive variance reduction for Monte Carlo integration by merging two expectations via common random numbers. The variance-reduction mechanism of the ARM estimator can also be attributed to either antithetic sampling in an augmented space, or the use of an optimal anti-symmetric "self-control" baseline function together with the REINFORCE estimator in that augmented space. Experimental results show the ARM estimator provides state-of-the-art performance in auto-encoding variational inference and maximum likelihood estimation, for discrete latent variable models with one or multiple stochastic binary layers. Python code for reproducible research is publicly available.

## Scalable Unbalanced Optimal Transport using Generative Adversarial Networks

- Karren D. Yang, Caroline Uhler
- abstract@open-review(Poster): Generative adversarial networks (GANs) are an expressive class of neural generative models with tremendous success in modeling high-dimensional continuous measures. In this paper, we present a scalable method for unbalanced optimal transport (OT) based on the generative-adversarial framework. We formulate unbalanced OT as a problem of simultaneously learning a transport map and a scaling factor that push a source measure to a target measure in a cost-optimal manner. We provide theoretical justification for this formulation, showing that it is closely related to an existing static formulation by Liero et al. (2018). We then propose an algorithm for solving this problem based on stochastic alternating gradient updates, similar in practice to GANs, and perform numerical experiments demonstrating how this methodology can be applied to population modeling.

## Unsupervised Discovery of Parts, Structure, and Dynamics

- Zhenjia Xu, *Zhijian Liu*, Chen Sun, Kevin Murphy, William T. Freeman, Joshua B. Tenenbaum, Jiajun Wu
- abstract@open-review(Poster): Humans easily recognize object parts and their hierarchical structure by watching how they move; they can then predict how each part moves in the future. In this paper, we propose a novel formulation that simultaneously learns a hierarchical, disentangled object representation and a dynamics model for object parts from unlabeled videos. Our Parts, Structure, and Dynamics (PSD) model learns to, first, recognize the object parts via a layered image representation; second, predict hierarchy via a structural descriptor that composes low-level concepts into a hierarchical structure; and third, model the system dynamics by predicting the future. Experiments on multiple real and synthetic datasets demonstrate that our PSD model works well on all three tasks: segmenting object parts, building their hierarchical structure, and capturing their motion distributions.

## Learning Multimodal Graph-to-Graph Translation for Molecule Optimization

- Wengong Jin, Kevin Yang, Regina Barzilay, Tommi Jaakkola
- abstract@open-review(Poster): We view molecule optimization as a graph-to-graph translation problem. The goal is to learn to map from one molecular graph to another with better properties based on an available corpus of paired molecules. Since molecules can be optimized in different ways, there are multiple viable translations for each input graph. A key challenge is therefore to model diverse translation outputs. Our primary contributions include a junction tree encoder-decoder for learning diverse graph translations along with a novel adversarial training method for aligning distributions of molecules. Diverse output distributions in our model are explicitly realized by low-dimensional latent vectors that modulate the translation process. We evaluate our model on multiple molecule optimization tasks and show that our model outperforms previous state-of-the-art baselines by a significant margin.

## Harmonizing Maximum Likelihood with GANs for Multimodal Conditional Generation

- Soochan Lee, Junsoo Ha, Gunhee Kim
- abstract@open-review(Poster): Recent advances in conditional image generation tasks, such as image-to-image translation and image inpainting, are largely accounted to the success of conditional GAN models, which are often optimized by the joint use of the GAN loss with the reconstruction loss. However, we reveal that this training recipe shared by almost all existing methods causes one critical side effect: lack of diversity in output samples. In order to accomplish both training stability and multimodal output generation, we propose novel training schemes with a new set of losses named moment reconstruction losses that simply replace the reconstruction loss. We show that our approach is applicable to any conditional generation tasks by performing thorough experiments on image-to-image translation, super-resolution and image inpainting using Cityscapes and CelebA dataset. Quantitative evaluations also confirm that our methods achieve a great diversity in outputs while retaining or even improving the visual fidelity of generated samples.

## Phase-Aware Speech Enhancement with Deep Complex U-Net

- Hyeong-Seok Choi, Jang-Hyun Kim, Jaesung Huh, Adrian Kim, Jung-Woo Ha, Kyogu Lee
- abstract@open-review(Poster): Most deep learning-based models for speech enhancement have mainly focused on estimating the magnitude of spectrogram while reusing the phase from noisy speech for reconstruction. This is due to the difficulty of estimating the phase of clean speech. To improve speech enhancement performance, we tackle the phase estimation problem in three ways. First, we propose Deep Complex U-Net, an advanced U-Net structured model incorporating well-defined complex-valued building blocks to deal with complex-valued spectrograms. Second, we propose a polar coordinate-wise complex-valued masking method to reflect the distribution of complex ideal ratio masks. Third, we define a novel loss function, weighted source-to-distortion ratio (wSDR) loss, which is designed to directly correlate with a quantitative evaluation measure. Our model was evaluated on a mixture of the Voice Bank corpus and DEMAND database, which has been widely used by many deep learning models for speech enhancement. Ablation experiments were conducted on the mixed dataset showing that all three proposed approaches are empirically valid. Experimental results show that the proposed method achieves state-of-the-art performance in all metrics, outperforming previous approaches by a large margin.

## The Comparative Power of ReLU Networks and Polynomial Kernels in the Presence of Sparse Latent Structure

- Frederic Koehler, Andrej Risteski
- abstract@open-review(Poster): There has been a large amount of interest, both in the past and particularly recently, into the relative advantage of different families of universal function approximators, for instance neural networks, polynomials, rational functions, etc. However, current research has focused almost exclusively on understanding this problem in a worst case setting: e.g. characterizing the best L1 or L_{infty} approximation in a box (or sometimes, even under an adversarially constructed data distribution.) In this setting many classical tools from approximation theory can be effectively used.

However, in typical applications we expect data to be high dimensional, but structured -- so, it would only be important to approximate the desired function well on the relevant part of its domain, e.g. a small manifold on which real input data actually lies. Moreover, even within this domain the desired quality of approximation may not be uniform; for instance in classification problems, the approximation needs to be more accurate near the decision boundary. These issues, to the best of our knowledge, have remain unexplored until now.

With this in mind, we analyze the performance of neural networks and polynomial kernels in a natural regression setting where the data enjoys sparse latent structure, and the labels depend in a simple way on the latent variables. We give an almost-tight theoretical analysis of the performance of both neural networks and polynomials for this problem, as well as verify our theory with simulations. Our results both involve new (complex-analytic) techniques, which may be of independent interest, and show substantial qualitative differences with what is known in the worst-case setting.

## Neural Probabilistic Motor Primitives for Humanoid Control

- Josh Merel, Leonard Hasenclever, Alexandre Galashov, Arun Ahuja, Vu Pham, Greg Wayne, Yee Whye Teh, Nicolas Heess
- abstract@open-review(Poster): We focus on the problem of learning a single motor module that can flexibly express a range of behaviors for the control of high-dimensional physically simulated humanoids. To do this, we propose a motor architecture that has the general structure of an inverse model with a latent-variable bottleneck. We show that it is possible to train this model entirely offline to compress thousands of expert policies and learn a motor primitive embedding space. The trained neural probabilistic motor primitive system can perform one-shot imitation of whole-body humanoid behaviors, robustly mimicking unseen trajectories. Additionally, we demonstrate that it is also straightforward to train controllers to reuse the learned motor primitive

space to solve tasks, and the resulting movements are relatively naturalistic. To support the training of our model, we compare two approaches for offline policy cloning, including an experience efficient method which we call linear feedback policy cloning. We encourage readers to view a supplementary video (https://youtu.be/CaDEf-QcKwA ) summarizing our results.

## Learning Two-layer Neural Networks with Symmetric Inputs

- Rong Ge, Rohith Kuditipudi, Zhize Li, Xiang Wang
- abstract@open-review(Poster): We give a new algorithm for learning a two-layer neural network under a very general class of input distributions. Assuming there is a ground-truth two-layer network y = A \sigma(Wx) + \xi, where A, W are weight matrices, \xi represents noise, and the number of neurons in the hidden layer is no larger than the input or output, our algorithm is guaranteed to recover the parameters A, W of the ground-truth network. The only requirement on the input x is that it is symmetric, which still allows highly complicated and structured input.

Our algorithm is based on the method-of-moments framework and extends several results in tensor decompositions. We use spectral algorithms to avoid the complicated non-convex optimization in learning neural networks. Experiments show that our algorithm can robustly learn the ground-truth neural network with a small number of samples for many symmetric input distributions.

## How Powerful are Graph Neural Networks?

- Keyulu Xu, *Weihua Hu*, Jure Leskovec, Stefanie Jegelka
- abstract@open-review(Oral): Graph Neural Networks (GNNs) are an effective framework for representation learning of graphs. GNNs follow a neighborhood aggregation scheme, where the representation vector of a node is computed by recursively aggregating and transforming representation vectors of its neighboring nodes. Many GNN variants have been proposed and have achieved state-of-the-art results on both node and graph classification tasks. However, despite GNNs revolutionizing graph representation learning, there is limited understanding of their representational properties and limitations. Here, we present a theoretical framework for analyzing the expressive power of GNNs to capture different graph structures. Our results characterize the discriminative power of popular GNN variants, such as Graph Convolutional Networks and GraphSAGE, and show that they cannot learn to distinguish certain simple graph structures. We then develop a simple architecture that is provably the most expressive among the class of GNNs and is as powerful as the Weisfeiler-Lehman graph isomorphism test. We empirically validate our theoretical findings on a number of graph classification benchmarks, and demonstrate that our model achieves state-of-the-art performance.

## Spectral Inference Networks: Unifying Deep and Spectral Learning

- David Pfau, Stig Petersen, Ashish Agarwal, David G. T. Barrett, Kimberly L. Stachenfeld
- abstract@open-review(Poster): We present Spectral Inference Networks, a framework for learning eigenfunctions of linear operators by stochastic optimization. Spectral Inference Networks generalize Slow Feature Analysis to generic symmetric operators, and are closely related to Variational Monte Carlo methods from computational physics. As such, they can be a powerful tool for unsupervised representation learning from video or graph-structured data. We cast training Spectral Inference Networks as a bilevel optimization problem, which allows for online learning of multiple eigenfunctions. We show results of training Spectral Inference Networks on problems in quantum mechanics and feature learning for videos on synthetic datasets. Our results demonstrate that Spectral Inference Networks accurately recover eigenfunctions of linear operators and can discover interpretable representations from video in a fully unsupervised manner.

## On Self Modulation for Generative Adversarial Networks

- Ting Chen, Mario Lucic, Neil Houlsby, Sylvain Gelly
- abstract@open-review(Poster): Training Generative Adversarial Networks (GANs) is notoriously challenging. We propose and study an architectural modification, self-modulation, which improves GAN performance across different data sets, architectures, losses, regularizers, and hyperparameter settings. Intuitively, self-modulation allows the intermediate feature maps of a generator to change as a function of the input noise vector. While reminiscent of other conditioning techniques, it requires no labeled data. In a large-scale empirical study we observe a relative decrease of 5%-35% in FID. Furthermore, all else being equal, adding this modification to the generator leads to improved performance in 124/144 (86%) of the studied settings. Self-modulation is a simple architectural change that requires no additional parameter tuning, which suggests that it can be applied readily to any GAN.

## AutoLoss: Learning Discrete Schedule for Alternate Optimization

- Haowen Xu, Hao Zhang, Zhiting Hu, Xiaodan Liang, Ruslan Salakhutdinov, Eric Xing
- abstract@open-review(Poster): Many machine learning problems involve iteratively and alternately optimizing different task objectives with respect to different sets of parameters. Appropriately scheduling the optimization of a task objective or a set of parameters is usually crucial to the quality of convergence. In this paper, we present AutoLoss, a meta-learning framework that automatically learns and determines the optimization schedule. AutoLoss provides a generic way to represent and learn the discrete optimization schedule from metadata, allows for a dynamic and data-driven schedule in ML problems that involve alternating updates of different parameters or from different loss objectives.

We apply AutoLoss on four ML tasks: d-ary quadratic regression, classification using a multi-layer perceptron (MLP), image generation using GANs, and multi-task neural machine translation (NMT). We show that the AutoLoss controller is able to capture the distribution of better optimization schedules that result in higher quality of convergence on all four tasks. The trained AutoLoss controller is generalizable -- it can guide and improve the learning of a new task model with different specifications, or on different datasets.

## Analyzing Inverse Problems with Invertible Neural Networks

- Lynton Ardizzone, Jakob Kruse, Carsten Rother, Ullrich Köthe
- abstract@open-review(Poster): For many applications, in particular in natural science, the task is to determine hidden system parameters from a set of measurements. Often, the forward process from parameter- to measurement-space is well-defined, whereas the inverse problem is ambiguous: multiple parameter sets can result in the same measurement. To fully characterize this ambiguity, the full posterior parameter distribution, conditioned on an observed measurement, has to be determined. We argue that a particular class of neural networks is well suited for this task – so-called Invertible Neural Networks (INNs). Unlike classical neural networks, which attempt to solve the ambiguous inverse problem directly, INNs focus on learning the forward process, using additional latent output variables to capture the information otherwise lost. Due to invertibility, a model of the corresponding inverse process is learned implicitly. Given a specific measurement and the distribution of the latent variables, the inverse pass of the INN provides the full posterior over parameter space. We prove theoretically and verify experimentally, on artificial data and real-world problems from medicine and astrophysics, that INNs are a powerful analysis tool to find multi-modalities in parameter space, uncover parameter correlations, and identify unrecoverable parameters.

## Learning Finite State Representations of Recurrent Policy Networks

- Anurag Koul, Alan Fern, Sam Greydanus

- abstract@open-review(Poster): Recurrent neural networks (RNNs) are an effective representation of control policies for a wide range of reinforcement and imitation learning problems. RNN policies, however, are particularly difficult to explain, understand, and analyze due to their use of continuous-valued memory vectors and observation features. In this paper, we introduce a new technique, Quantized Bottleneck Insertion, to learn finite representations of these vectors and features. The result is a quantized representation of the RNN that can be analyzed to improve our understanding of memory use and general behavior. We present results of this approach on synthetic environments and six Atari games. The resulting finite representations are surprisingly small in some cases, using as few as 3 discrete memory states and 10 observations for a perfect Pong policy. We also show that these finite policy representations lead to improved interpretability.

## Measuring and regularizing networks in function space

- Ari Benjamin, David Rolnick, Konrad Kording
- abstract@open-review(Poster): To optimize a neural network one often thinks of optimizing its parameters, but it is ultimately a matter of optimizing the function that maps inputs to outputs. Since a change in the parameters might serve as a poor proxy for the change in the function, it is of some concern that primacy is given to parameters but that the correspondence has not been tested. Here, we show that it is simple and computationally feasible to calculate distances between functions in a $L^2$ Hilbert space. We examine how typical networks behave in this space, and compare how parameter $\ell^2$ distances compare to function $L^2$ distances between various points of an optimization trajectory. We find that the two distances are nontrivially related. In particular, the $L^2/\ell^2$ ratio decreases throughout optimization, reaching a steady value around when test error plateaus. We then investigate how the $L^2$ distance could be applied directly to optimization. We first propose that in multitask learning, one can avoid catastrophic forgetting by directly limiting how much the input/output function changes between tasks. Secondly, we propose a new learning rule that constrains the distance a network can travel through $L^2$-space in any one update. This allows new examples to be learned in a way that minimally interferes with what has previously been learned. These applications demonstrate how one can measure and regularize function distances directly, without relying on parameters or local approximations like loss curvature.

## No Training Required: Exploring Random Encoders for Sentence Classification

- John Wieting, Douwe Kiela
- abstract@open-review(Poster): We explore various methods for computing sentence representations from pre-trained word embeddings without any training, i.e., using nothing but random parameterizations. Our aim is to put sentence embeddings on more solid footing by 1) looking at how much modern sentence embeddings gain over random methods---as it turns out, surprisingly little; and by 2) providing the field with more appropriate baselines going forward---which are, as it turns out, quite strong. We also make important observations about proper experimental protocol for sentence classification evaluation, together with recommendations for future research.

## Deep Frank-Wolfe For Neural Network Optimization

- Leonard Berrada, Andrew Zisserman, M. Pawan Kumar
- abstract@open-review(Poster): Learning a deep neural network requires solving a challenging optimization problem: it is a high-dimensional, non-convex and non-smooth minimization problem with a large number of terms. The current practice in neural network optimization is to rely on the stochastic gradient descent (SGD) algorithm or its adaptive variants. However, SGD requires a hand-designed schedule for the learning rate. In addition, its adaptive variants tend to produce solutions that generalize less well on unseen data than SGD with a hand-designed schedule. We present an optimization method that offers empirically the best of both worlds: our algorithm yields good generalization performance while requiring only one hyper-parameter. Our approach is based on a composite proximal framework, which exploits the compositional nature of deep neural networks and can leverage powerful convex optimization algorithms by design. Specifically, we employ the Frank-Wolfe (FW) algorithm for SVM, which computes an optimal step-size in closed-form at each time-step. We further show that the descent direction is given by a simple backward pass in the network, yielding the same computational cost per iteration as SGD. We present experiments on the CIFAR and SNLI data sets, where we demonstrate the significant superiority of our method over Adam, Adagrad, as well as the recently proposed BPGrad and AMSGrad. Furthermore, we compare our algorithm to SGD with a hand-designed learning rate schedule, and show that it provides similar generalization while often converging faster. The code is publicly available at https://github.com/oval-group/dfw.

## Quasi-hyperbolic momentum and Adam for deep learning

- Jerry Ma, Denis Yarats
- abstract@open-review(Poster): Momentum-based acceleration of stochastic gradient descent (SGD) is widely used in deep learning. We propose the quasi-hyperbolic momentum algorithm (QHM) as an extremely simple alteration of momentum SGD, averaging a plain SGD step with a momentum step. We describe numerous connections to and identities with other algorithms, and we characterize the set of two-state optimization algorithms that QHM can recover. Finally, we propose a QH variant of Adam called QHAdam, and we empirically demonstrate that our algorithms lead to significantly improved training in a variety of settings, including a new state-of-the-art result on WMT16 EN-DE. We hope that these empirical results, combined with the conceptual and practical simplicity of QHM and QHAdam, will spur interest from both practitioners and researchers. Code is immediately available.

## On Computation and Generalization of Generative Adversarial Networks under Spectrum Control

- Haoming Jiang, Zhehui Chen, Minshuo Chen, Feng Liu, Dingding Wang, Tuo Zhao
- abstract@open-review(Poster): Generative Adversarial Networks (GANs), though powerful, is hard to train. Several recent works (Brock et al., 2016; Miyato et al., 2018) suggest that controlling the spectra of weight matrices in the discriminator can significantly improve the training of GANs. Motivated by their discovery, we propose a new framework for training GANs, which allows more flexible spectrum control (e.g., making the weight matrices of the discriminator have slow singular value decays). Specifically, we propose a new reparameterization approach for the weight matrices of the discriminator in GANs, which allows us to directly manipulate the spectra of the weight matrices through various regularizers and constraints, without intensively computing singular value decompositions. Theoretically, we further show that the spectrum control improves the generalization ability of GANs. Our experiments on CIFAR-10, STL-10, and ImgaeNet datasets confirm that compared to other competitors, our proposed method is capable of generating images with better or equal quality by utilizing spectral normalization and encouraging the slow singular value decay.

## Big-Little Net: An Efficient Multi-Scale Feature Representation for Visual and Speech Recognition

- Chun-Fu (Richard) Chen, Quanfu Fan, Neil Mallinar, Tom Sercu, Rogerio Feris
- abstract@open-review(Poster): In this paper, we propose a novel Convolutional Neural Network (CNN) architecture for learning multi-scale feature representations with good tradeoffs between speed and accuracy. This is achieved by using a multi-branch network, which has different computational complexity at different branches with different resolutions. Through frequent merging of features from branches at distinct scales, our model obtains multi-scale features while using less computation. The proposed approach demonstrates improvement of model efficiency and performance on both object recognition and speech recognition tasks, using popular architectures including ResNet, ResNeXt and SEResNeXt. For object recognition, our approach reduces computation by 1/3 while improving accuracy significantly over 1% point than the baselines, and the computational savings can be higher up to 1/2 without compromising the accuracy. Our model also surpasses state-of-the-art CNN acceleration approaches by a large margin in terms of accuracy and FLOPs. On the task of speech recognition, our proposed multi-scale CNNs save 30% FLOPs with slightly better word error rates, showing good generalization across domains.

## Deep Lagrangian Networks: Using Physics as Model Prior for Deep Learning

- Michael Lutter, Christian Ritter, Jan Peters
- abstract@open-review(Poster): Deep learning has achieved astonishing results on many tasks with large amounts of data and generalization within the proximity of training data. For many important real-world applications, these requirements are unfeasible and additional prior knowledge on the task domain is required to overcome the resulting problems. In particular, learning physics models for model-based control requires robust extrapolation from fewer samples – often collected online in real-time – and model errors may lead to drastic damages of the system. Directly incorporating physical insight has enabled us to obtain a novel deep model learning approach that extrapolates well while requiring fewer samples. As a first example, we propose Deep Lagrangian Networks (DeLaN) as a deep network structure upon which Lagrangian Mechanics have been imposed. DeLaN can learn the equations of motion of a mechanical system (i.e., system dynamics) with a deep network efficiently while ensuring physical plausibility. The resulting DeLaN network performs very well at robot tracking control. The proposed method did not only outperform previous model learning approaches at learning speed but exhibits substantially improved and more robust extrapolation to novel trajectories and learns online in real-time.

## Adversarial Audio Synthesis

- Chris Donahue, Julian McAuley, Miller Puckette
- abstract@open-review(Poster): Audio signals are sampled at high temporal resolutions, and learning to synthesize audio requires capturing structure across a range of timescales. Generative adversarial networks (GANs) have seen wide success at generating images that are both locally and globally coherent, but they have seen little application to audio generation. In this paper we introduce WaveGAN, a first attempt at applying GANs to unsupervised synthesis of raw-waveform audio. WaveGAN is capable of synthesizing one second slices of audio waveforms with global coherence, suitable for sound effect generation. Our experiments demonstrate that—without labels—WaveGAN learns to produce intelligible words when trained on a small-vocabulary speech dataset, and can also synthesize audio from other domains such as drums, bird vocalizations, and piano. We compare WaveGAN to a method which applies GANs designed for image generation on image-like audio feature representations, finding both approaches to be promising.

## A Data-Driven and Distributed Approach to Sparse Signal Representation and Recovery

- Ali Mousavi, Gautam Dasarathy, Richard G. Baraniuk
- abstract@open-review(Poster): In this paper, we focus on two challenges which offset the promise of sparse signal representation, sensing, and recovery. First, real-world signals can seldom be described as perfectly sparse vectors in a known basis, and traditionally used random measurement schemes are seldom optimal for sensing them. Second, existing signal recovery algorithms are usually not fast enough to make them applicable to real-time problems. In this paper, we address these two challenges by presenting a novel framework based on deep learning. For the first challenge, we cast the problem of finding informative measurements by using a maximum likelihood (ML) formulation and show how we can build a data-driven dimensionality reduction protocol for sensing signals using convolutional architectures. For the second challenge, we discuss and analyze a novel parallelization scheme and show it significantly speeds-up the signal recovery process. We demonstrate the significant improvement our method obtains over competing methods through a series of experiments.

## The Laplacian in RL: Learning Representations with Efficient Approximations

- Yifan Wu, George Tucker, Ofir Nachum
- abstract@open-review(Poster): The smallest eigenvectors of the graph Laplacian are well-known to provide a succinct representation of the geometry of a weighted graph. In reinforcement learning (RL), where the weighted graph may be interpreted as the state transition process induced by a behavior policy acting on the environment, approximating the eigenvectors of the Laplacian provides a promising approach to state representation learning. However, existing methods for performing this approximation are ill-suited in general RL settings for two main reasons: First, they are computationally expensive, often requiring operations on large matrices. Second, these methods lack adequate justification beyond simple, tabular, finite-state settings. In this paper, we present a fully general and scalable method for approximating the eigenvectors of the Laplacian in a model-free RL context. We systematically evaluate our approach and empirically show that it generalizes beyond the tabular, finite-state setting. Even in tabular, finite-state settings, its ability to approximate the eigenvectors outperforms previous proposals. Finally, we show the potential benefits of using a Laplacian representation learned using our method in goal-achieving RL tasks, providing evidence that our technique can be used to significantly improve the performance of an RL agent.

## On the Relation Between the Sharpest Directions of DNN Loss and the SGD Step Length

- Stanisław Jastrzębski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, Amos Storkey
- abstract@open-review(Poster): The training of deep neural networks with Stochastic Gradient Descent (SGD) with a large learning rate or a small batch-size typically ends in flat regions of the weight space, as indicated by small eigenvalues of the Hessian of the training loss. This was found to correlate with a good final generalization performance. In this paper we extend previous work by investigating the curvature of the loss surface along the whole training trajectory, rather than only at the endpoint. We find that initially SGD visits increasingly sharp regions, reaching a maximum sharpness determined by both the learning rate and the batch-size of SGD. At this peak value SGD starts to fail to minimize the loss along directions in the loss surface corresponding to the largest curvature (sharpest directions). To further investigate the effect of these dynamics in the training process, we study a variant of SGD using a reduced learning rate along the sharpest directions which we show can improve training speed while finding both sharper and better generalizing solution, compared to vanilla SGD. Overall, our results show that the SGD dynamics in the subspace of the sharpest directions influence the regions that SGD steers to (where larger learning rate or smaller batch size result in wider regions visited), the overall training speed, and the generalization ability of the final model.

## Discriminator-Actor-Critic: Addressing Sample Inefficiency and Reward Bias in Adversarial Imitation Learning

- Ilya Kostrikov, Kumar Krishna Agrawal, Debidatta Dwibedi, Sergey Levine, Jonathan Tompson
- abstract@open-review(Poster): We identify two issues with the family of algorithms based on the Adversarial Imitation Learning framework. The first problem is implicit bias present in the reward functions used in these algorithms. While these biases might work well for some environments, they can also lead to sub-optimal behavior in others. Secondly, even though these algorithms can learn from few expert demonstrations, they require a prohibitively large number of interactions with the environment in order to imitate the expert for many real-world applications. In order to address these issues, we propose a new algorithm called Discriminator-Actor-Critic that uses off-policy Reinforcement Learning to reduce policy-environment interaction sample complexity by an average factor of 10. Furthermore, since our reward function is designed to be unbiased, we can apply our algorithm to many problems without making any task-specific adjustments.

## GENERATING HIGH FIDELITY IMAGES WITH SUBSCALE PIXEL NETWORKS AND MULTIDIMENSIONAL UPSCALING

- Jacob Menick, Nal Kalchbrenner
- abstract@open-review(Oral): The unconditional generation of high fidelity images is a longstanding benchmark for testing the performance of image decoders. Autoregressive image models have been able to generate small images unconditionally, but the extension of these methods to large images where fidelity can be more readily assessed has remained an open problem. Among the major challenges are the capacity to encode the vast previous

context and the sheer difficulty of learning a distribution that preserves both global semantic coherence and exactness of detail. To address the former challenge, we propose the Subscale Pixel Network (SPN), a conditional decoder architecture that generates an image as a sequence of image slices of equal size. The SPN compactly captures image-wide spatial dependencies and requires a fraction of the memory and the computation. To address the latter challenge, we propose to use multidimensional upscaling to grow an image in both size and depth via intermediate stages corresponding to distinct SPNs. We evaluate SPNs on the unconditional generation of CelebAHQ of size 256 and of ImageNet from size 32 to 128. We achieve state-of-the-art likelihood results in multiple settings, set up new benchmark results in previously unexplored settings and are able to generate very high fidelity large scale samples on the basis of both datasets.

## Excessive Invariance Causes Adversarial Vulnerability

- Joern-Henrik Jacobsen, Jens Behrmann, Richard Zemel, Matthias Bethge
- abstract@open-review(Poster): Despite their impressive performance, deep neural networks exhibit striking failures on out-of-distribution inputs. One core idea of adversarial example research is to reveal neural network errors under such distribution shifts. We decompose these errors into two complementary sources: sensitivity and invariance. We show deep networks are not only too sensitive to task-irrelevant changes of their input, as is well-known from epsilon-adversarial examples, but are also too invariant to a wide range of task-relevant changes, thus making vast regions in input space vulnerable to adversarial attacks. We show such excessive invariance occurs across various tasks and architecture types. On MNIST and ImageNet one can manipulate the class-specific content of almost any image without changing the hidden activations. We identify an insufficiency of the standard cross-entropy loss as a reason for these failures. Further, we extend this objective based on an information-theoretic analysis so it encourages the model to consider all task-dependent features in its decision. This provides the first approach tailored explicitly to overcome excessive invariance and resulting vulnerabilities.

## Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality

- Taiji Suzuki
- abstract@open-review(Poster): Deep learning has shown high performances in various types of tasks from visual recognition to natural language processing, which indicates superior flexibility and adaptivity of deep learning. To understand this phenomenon theoretically, we develop a new approximation and estimation error analysis of deep learning with the ReLU activation for functions in a Besov space and its variant with mixed smoothness. The Besov space is a considerably general function space including the Holder space and Sobolev space, and especially can capture spatial inhomogeneity of smoothness. Through the analysis in the Besov space, it is shown that deep learning can achieve the minimax optimal rate and outperform any non-adaptive (linear) estimator such as kernel ridge regression, which shows that deep learning has higher adaptivity to the spatial inhomogeneity of the target function than other estimators such as linear ones. In addition to this, it is shown that deep learning can avoid the curse of dimensionality if the target function is in a mixed smooth Besov space. We also show that the dependency of the convergence rate on the dimensionality is tight due to its minimax optimality. These results support high adaptivity of deep learning and its superior ability as a feature extractor.

## AntisymmetricRNN: A Dynamical System View on Recurrent Neural Networks

- Bo Chang, Minmin Chen, Eldad Haber, Ed H. Chi
- abstract@open-review(Poster): Recurrent neural networks have gained widespread use in modeling sequential data. Learning long-term dependencies using these models remains difficult though, due to exploding or vanishing gradients. In this paper, we draw connections between recurrent networks and ordinary differential equations. A special form of recurrent networks called the AntisymmetricRNN is proposed under this theoretical framework, which is able to capture long-term dependencies thanks to the stability property of its underlying differential equation. Existing approaches to improving RNN trainability often incur significant computation overhead. In comparison, AntisymmetricRNN achieves the same goal by design. We showcase the advantage of this new architecture through extensive simulations and experiments. AntisymmetricRNN exhibits much more predictable dynamics. It outperforms regular LSTM models on tasks requiring long-term memory and matches the performance on tasks where short-term dependencies dominate despite being much simpler.

## Discriminator Rejection Sampling

- Samaneh Azadi, Catherine Olsson, Trevor Darrell, Ian Goodfellow, Augustus Odena
- abstract@open-review(Poster): We propose a rejection sampling scheme using the discriminator of a GAN to approximately correct errors in the GAN generator distribution. We show that under quite strict assumptions, this will allow us to recover the data distribution exactly. We then examine where those strict assumptions break down and design a practical algorithm—called Discriminator Rejection Sampling (DRS)—that can be used on real data-sets. Finally, we demonstrate the efficacy of DRS on a mixture of Gaussians and on the state of the art SAGAN model. On ImageNet, we train an improved baseline that increases the best published Inception Score from 52.52 to 62.36 and reduces the Frechet Inception Distance from 18.65 to 14.79. We then use DRS to further improve on this baseline, improving the Inception Score to 76.08 and the FID to 13.75.

## Unsupervised Learning via Meta-Learning

- Kyle Hsu, Sergey Levine, Chelsea Finn
- abstract@open-review(Poster): A central goal of unsupervised learning is to acquire representations from unlabeled data or experience that can be used for more effective learning of downstream tasks from modest amounts of labeled data. Many prior unsupervised learning works aim to do so by developing proxy objectives based on reconstruction, disentanglement, prediction, and other metrics. Instead, we develop an unsupervised meta-learning method that explicitly optimizes for the ability to learn a variety of tasks from small amounts of data. To do so, we construct tasks from unlabeled data in an automatic way and run meta-learning over the constructed tasks. Surprisingly, we find that, when integrated with meta-learning, relatively simple task construction mechanisms, such as clustering embeddings, lead to good performance on a variety of downstream, human-specified tasks. Our experiments across four image datasets indicate that our unsupervised meta-learning approach acquires a learning algorithm without any labeled data that is applicable to a wide range of downstream classification tasks, improving upon the embedding learned by four prior unsupervised learning methods.

## Recurrent Experience Replay in Distributed Reinforcement Learning

- Steven Kapturowski, Georg Ostrovski, John Quan, Remi Munos, Will Dabney
- abstract@open-review(Poster): Building on the recent successes of distributed training of RL agents, in this paper we investigate the training of RNN-based RL agents from distributed prioritized experience replay. We study the effects of parameter lag resulting in representational drift and recurrent state staleness and empirically derive an improved training strategy. Using a single network architecture and fixed set of hyper-parameters, the resulting agent, Recurrent Replay Distributed DQN, quadruples the previous state of the art on Atari-57, and matches the state of the art on DMLab-30. It is the first agent to exceed human-level performance in 52 of the 57 Atari games.

## Query-Efficient Hard-label Black-box Attack: An Optimization-based Approach

- Minhao Cheng, Thong Le, Pin-Yu Chen, Huan Zhang, JinFeng Yi, Cho-Jui Hsieh

- abstract@open-review(Poster): We study the problem of attacking machine learning models in the hard-label black-box setting, where no model information is revealed except that the attacker can make queries to probe the corresponding hard-label decisions. This is a very challenging problem since the direct extension of state-of-the-art white-box attacks (e.g., C&W or PGD) to the hard-label black-box setting will require minimizing a non-continuous step function, which is combinatorial and cannot be solved by a gradient-based optimizer. The only two current approaches are based on random walk on the boundary (Brendel et al., 2017) and random trials to evaluate the loss function (Ilyas et al., 2018), which require lots of queries and lacks convergence guarantees. We propose a novel way to formulate the hard-label black-box attack as a real-valued optimization problem which is usually continuous and can be solved by any zeroth order optimization algorithm. For example, using the Randomized Gradient-Free method (Nesterov & Spokoiny, 2017), we are able to bound the number of iterations needed for our algorithm to achieve stationary points under mild assumptions. We demonstrate that our proposed method outperforms the previous stochastic approaches to attacking convolutional neural networks on MNIST, CIFAR, and ImageNet datasets. More interestingly, we show that the proposed algorithm can also be used to attack other discrete and non-continuous machine learning models, such as Gradient Boosting Decision Trees (GBDT).

## Multi-class classification without multi-class labels

- Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, Zsolt Kira
- abstract@open-review(Poster): This work presents a new strategy for multi-class classification that requires no class-specific labels, but instead leverages pairwise similarity between examples, which is a weaker form of annotation. The proposed method, meta classification learning, optimizes a binary classifier for pairwise similarity prediction and through this process learns a multi-class classifier as a submodule. We formulate this approach, present a probabilistic graphical model for it, and derive a surprisingly simple loss function that can be used to learn neural network-based models. We then demonstrate that this same framework generalizes to the supervised, unsupervised cross-task, and semi-supervised settings. Our method is evaluated against state of the art in all three learning paradigms and shows a superior or comparable accuracy, providing evidence that learning multi-class classification without multi-class labels is a viable learning option.

## Large-Scale Answerer in Questioner's Mind for Visual Dialog Question Generation

- Sang-Woo Lee, Tong Gao, Sohee Yang, Jaejun Yoo, Jung-Woo Ha
- abstract@open-review(Poster): Answerer in Questioner's Mind (AQM) is an information-theoretic framework that has been recently proposed for task-oriented dialog systems. AQM benefits from asking a question that would maximize the information gain when it is asked. However, due to its intrinsic nature of explicitly calculating the information gain, AQM has a limitation when the solution space is very large. To address this, we propose AQM+ that can deal with a large-scale problem and ask a question that is more coherent to the current context of the dialog. We evaluate our method on GuessWhich, a challenging task-oriented visual dialog problem, where the number of candidate classes is near 10K. Our experimental results and ablation studies show that AQM+ outperforms the state-of-the-art models by a remarkable margin with a reasonable approximation. In particular, the proposed AQM+ reduces more than 60% of error as the dialog proceeds, while the comparative algorithms diminish the error by less than 6%. Based on our results, we argue that AQM+ is a general task-oriented dialog algorithm that can be applied for non-yes-or-no responses.

## Unsupervised Hyper-alignment for Multilingual Word Embeddings

- Jean Alaux, Edouard Grave, Marco Cuturi, Armand Joulin
- abstract@open-review(Poster): We consider the problem of aligning continuous word representations, learned in multiple languages, to a common space. It was recently shown that, in the case of two languages, it is possible to learn such a mapping without supervision. This paper extends this line of work to the problem of aligning multiple languages to a common space. A solution is to independently map all languages to a pivot language. Unfortunately, this degrades the quality of indirect word translation. We thus propose a novel formulation that ensures composable mappings, leading to better alignments. We evaluate our method by jointly aligning word vectors in eleven languages, showing consistent improvement with indirect mappings while maintaining competitive performance on direct word translation.

## Diversity is All You Need: Learning Skills without a Reward Function

- Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, Sergey Levine
- abstract@open-review(Poster): Intelligent creatures can explore their environments and learn useful skills without supervision. In this paper, we propose ``Diversity is All You Need"(DIAYN), a method for learning useful skills without a reward function. Our proposed method learns skills by maximizing an information theoretic objective using a maximum entropy policy. On a variety of simulated robotic tasks, we show that this simple objective results in the unsupervised emergence of diverse skills, such as walking and jumping. In a number of reinforcement learning benchmark environments, our method is able to learn a skill that solves the benchmark task despite never receiving the true task reward. We show how pretrained skills can provide a good parameter initialization for downstream tasks, and can be composed hierarchically to solve complex, sparse reward tasks. Our results suggest that unsupervised discovery of skills can serve as an effective pretraining mechanism for overcoming challenges of exploration and data efficiency in reinforcement learning.

## Solving the Rubik's Cube with Approximate Policy Iteration

- Stephen McAleer, Forest Agostinelli, Alexander Shmakov, Pierre Baldi
- abstract@open-review(Poster): Recently, Approximate Policy Iteration (API) algorithms have achieved super-human proficiency in two-player zero-sum games such as Go, Chess, and Shogi without human data. These API algorithms iterate between two policies: a slow policy (tree search), and a fast policy (a neural network). In these two-player games, a reward is always received at the end of the game. However, the Rubik’s Cube has only a single solved state, and episodes are not guaranteed to terminate. This poses a major problem for these API algorithms since they rely on the reward received at the end of the game. We introduce Autodidactic Iteration: an API algorithm that overcomes the problem of sparse rewards by training on a distribution of states that allows the reward to propagate from the goal state to states farther away. Autodidactic Iteration is able to learn how to solve the Rubik’s Cube and the 15-puzzle without relying on human data. Our algorithm is able to solve 100% of randomly scrambled cubes while achieving a median solve length of 30 moves — less than or equal to solvers that employ human domain knowledge.

## Dynamic Channel Pruning: Feature Boosting and Suppression

- Xitong Gao, Yiren Zhao, Łukasz Dudziak, Robert Mullins, Cheng-zhong Xu
- abstract@open-review(Poster): Making deep convolutional neural networks more accurate typically comes at the cost of increased computational and memory resources. In this paper, we reduce this cost by exploiting the fact that the importance of features computed by convolutional layers is highly input-dependent, and propose feature boosting and suppression (FBS), a new method to predictively amplify salient convolutional channels and skip unimportant ones at run-time. FBS introduces small auxiliary connections to existing convolutional layers. In contrast to channel pruning methods which permanently remove channels, it preserves the full network structures and accelerates convolution by dynamically skipping unimportant input and output channels. FBS-augmented networks are trained with conventional stochastic gradient descent, making it readily available for many state-of-the-art CNNs. We compare FBS to a range of existing channel pruning and dynamic execution schemes and demonstrate large improvements on ImageNet classification. Experiments show that FBS can respectively provide 5× and 2× savings in compute on VGG-16 and ResNet-18, both with less than 0.6% top-5 accuracy loss.

## Beyond Pixel Norm-Balls: Parametric Adversaries using an Analytically Differentiable Renderer

- Hsueh-Ti Derek Liu, Michael Tao, Chun-Liang Li, Derek Nowrouzezahrai, Alec Jacobson
- abstract@open-review(Poster): Many machine learning image classifiers are vulnerable to adversarial attacks, inputs with perturbations designed to intentionally trigger misclassification. Current adversarial methods directly alter pixel colors and evaluate against pixel norm-balls: pixel perturbations smaller than a specified magnitude, according to a measurement norm. This evaluation, however, has limited practical utility since perturbations in the pixel space do not correspond to underlying real-world phenomena of image formation that lead to them and has no security motivation attached. Pixels in natural images are measurements of light that has interacted with the geometry of a physical scene. As such, we propose a novel evaluation measure, parametric norm-balls, by directly perturbing physical parameters that underly image formation. One enabling contribution we present is a physically-based differentiable renderer that allows us to propagate pixel gradients to the parametric space of lighting and geometry. Our approach enables physically-based adversarial attacks, and our differentiable renderer leverages models from the interactive rendering literature to balance the performance and accuracy trade-offs necessary for a memory-efficient and scalable adversarial data augmentation workflow.

## Deterministic PAC-Bayesian generalization bounds for deep networks via generalizing noise-resilience

- Vaishnavh Nagarajan, Zico Kolter
- abstract@open-review(Poster): The ability of overparameterized deep networks to generalize well has been linked to the fact that stochastic gradient descent (SGD) finds solutions that lie in flat, wide minima in the training loss -- minima where the output of the network is resilient to small random noise added to its parameters. So far this observation has been used to provide generalization guarantees only for neural networks whose parameters are either \textit{stochastic} or \textit{compressed}. In this work, we present a general PAC-Bayesian framework that leverages this observation to provide a bound on the original network learned -- a network that is deterministic and uncompressed. What enables us to do this is a key novelty in our approach: our framework allows us to show that if on training data, the interactions between the weight matrices satisfy certain conditions that imply a wide training loss minimum, these conditions themselves {\em generalize} to the interactions between the matrices on test data, thereby implying a wide test loss minimum. We then apply our general framework in a setup where we assume that the pre-activation values of the network are not too small (although we assume this only on the training data). In this setup, we provide a generalization guarantee for the original (deterministic, uncompressed) network, that does not scale with product of the spectral norms of the weight matrices -- a guarantee that would not have been possible with prior approaches.

## Learning-Based Frequency Estimation Algorithms

- Chen-Yu Hsu, Piotr Indyk, Dina Katabi, Ali Vakilian
- abstract@open-review(Poster): Estimating the frequencies of elements in a data stream is a fundamental task in data analysis and machine learning. The problem is typically addressed using streaming algorithms which can process very large data using limited storage. Today's streaming algorithms, however, cannot exploit patterns in their input to improve performance. We propose a new class of algorithms that automatically learn relevant patterns in the input data and use them to improve its frequency estimates. The proposed algorithms combine the benefits of machine learning with the formal guarantees available through algorithm theory. We prove that our learning-based algorithms have lower estimation errors than their non-learning counterparts. We also evaluate our algorithms on two real-world datasets and demonstrate empirically their performance gains.

## A Unified Theory of Early Visual Representations from Retina to Cortex through Anatomically Constrained Deep CNNs

- Jack Lindsey, Samuel A. Ocko, Surya Ganguli, Stephane Deny
- abstract@open-review(Oral): The vertebrate visual system is hierarchically organized to process visual information in successive stages. Neural representations vary drastically across the first stages of visual processing: at the output of the retina, ganglion cell receptive fields (RFs) exhibit a clear antagonistic center-surround structure, whereas in the primary visual cortex (V1), typical RFs are sharply tuned to a precise orientation. There is currently no unified theory explaining these differences in representations across layers. Here, using a deep convolutional neural network trained on image recognition as a model of the visual system, we show that such differences in representation can emerge as a direct consequence of different neural resource constraints on the retinal and cortical networks, and for the first time we find a single model from which both geometries spontaneously emerge at the appropriate stages of visual processing. The key constraint is a reduced number of neurons at the retinal output, consistent with the anatomy of the optic nerve as a stringent bottleneck. Second, we find that, for simple downstream cortical networks, visual representations at the retinal output emerge as nonlinear and lossy feature detectors, whereas they emerge as linear and faithful encoders of the visual scene for more complex cortical networks. This result predicts that the retinas of small vertebrates (e.g. salamander, frog) should perform sophisticated nonlinear computations, extracting features directly relevant to behavior, whereas retinas of large animals such as primates should mostly encode the visual scene linearly and respond to a much broader range of stimuli. These predictions could reconcile the two seemingly incompatible views of the retina as either performing feature extraction or efficient coding of natural scenes, by suggesting that all vertebrates lie on a spectrum between these two objectives, depending on the degree of neural resources allocated to their visual system.

## From Hard to Soft: Understanding Deep Network Nonlinearities via Vector Quantization and Statistical Inference

- Randall Balestriero, Richard Baraniuk
- abstract@open-review(Poster): Nonlinearity is crucial to the performance of a deep (neural) network (DN). To date there has been little progress understanding the menagerie of available nonlinearities, but recently progress has been made on understanding the r\^{o}le played by piecewise affine and convex nonlinearities like the ReLU and absolute value activation functions and max-pooling. In particular, DN layers constructed from these operations can be interpreted as {\em max-affine spline operators} (MASOs) that have an elegant link to vector quantization (VQ) and $K$-means. While this is good theoretical progress, the entire MASO approach is predicated on the requirement that the nonlinearities be piecewise affine and convex, which precludes important activation functions like the sigmoid, hyperbolic tangent, and softmax. {\em This paper extends the MASO framework to these and an infinitely large class of new nonlinearities by linking deterministic MASOs with probabilistic Gaussian Mixture Models (GMMs).} We show that, under a GMM, piecewise affine, convex nonlinearities like ReLU, absolute value, and max-pooling can be interpreted as solutions to certain natural `hard'' VQ inference problems, while sigmoid, hyperbolic tangent, and softmax can be interpreted as solutions to correspondingsoft'' VQ inference problems. We further extend the framework by hybridizing the hard and soft VQ optimizations to create a $\beta$-VQ inference that interpolates between hard, soft, and linear VQ inference. A prime example of a $\beta$-VQ DN nonlinearity is the {\em swish} nonlinearity, which offers state-of-the-art performance in a range of computer vision tasks but was developed ad hoc by experimentation. Finally, we validate with experiments an important assertion of our theory, namely that DN performance can be significantly improved by enforcing orthogonality in its linear filters.

## Episodic Curiosity through Reachability

- Nikolay Savinov, Anton Raichuk, Damien Vincent, Raphael Marinier, Marc Pollefeys, Timothy Lillicrap, Sylvain Gelly
- abstract@open-review(Poster): Rewards are sparse in the real world and most of today's reinforcement learning algorithms struggle with such sparsity. One solution to this problem is to allow the agent to create rewards for itself - thus making rewards dense and more suitable for learning. In particular, inspired by curious behaviour in animals, observing something novel could be rewarded with a bonus. Such bonus is summed up with the real task reward - making it possible for RL algorithms to learn from the combined reward. We propose a new curiosity method which uses episodic memory to form the novelty bonus. To determine the bonus, the current observation is compared with the observations in memory. Crucially, the comparison is done based on how many environment steps it takes to reach the current observation from those in memory - which incorporates rich information about environment dynamics. This allows us to overcome the known "couch-potato" issues of prior work - when the agent finds a way to instantly gratify itself by exploiting

actions which lead to hardly predictable consequences. We test our approach in visually rich 3D environments in ViZDoom, DMLab and MuJoCo. In navigational tasks from ViZDoom and DMLab, our agent outperforms the state-of-the-art curiosity method ICM. In MuJoCo, an ant equipped with our curiosity module learns locomotion out of the first-person-view curiosity only. The code is available at https://github.com/google-research/episodic-curiosity/.

## Bayesian Prediction of Future Street Scenes using Synthetic Likelihoods

- Apratim Bhattacharyya, Mario Fritz, Bernt Schiele
- abstract@open-review(Poster): For autonomous agents to successfully operate in the real world, the ability to anticipate future scene states is a key competence. In real-world scenarios, future states become increasingly uncertain and multi-modal, particularly on long time horizons. Dropout based Bayesian inference provides a computationally tractable, theoretically well grounded approach to learn different hypotheses/models to deal with uncertain futures and make predictions that correspond well to observations -- are well calibrated. However, it turns out that such approaches fall short to capture complex real-world scenes, even falling behind in accuracy when compared to the plain deterministic approaches. This is because the used log-likelihood estimate discourages diversity. In this work, we propose a novel Bayesian formulation for anticipating future scene states which leverages synthetic likelihoods that encourage the learning of diverse models to accurately capture the multi-modal nature of future scene states. We show that our approach achieves accurate state-of-the-art predictions and calibrated probabilities through extensive experiments for scene anticipation on Cityscapes dataset. Moreover, we show that our approach generalizes across diverse tasks such as digit generation and precipitation forecasting.

## Gradient Descent Provably Optimizes Over-parameterized Neural Networks

- Simon S. Du, Xiyu Zhai, Barnabas Poczos, Aarti Singh
- abstract@open-review(Poster): One of the mysteries in the success of neural networks is randomly initialized first order methods like gradient descent can achieve zero training loss even though the objective function is non-convex and non-smooth. This paper demystifies this surprising phenomenon for two-layer fully connected ReLU activated neural networks. For an $m$ hidden node shallow neural network with ReLU activation and $n$ training data, we show as long as $m$ is large enough and no two inputs are parallel, randomly initialized gradient descent converges to a globally optimal solution at a linear convergence rate for the quadratic loss function.

Our analysis relies on the following observation: over-parameterization and random initialization jointly restrict every weight vector to be close to its initialization for all iterations, which allows us to exploit a strong convexity-like property to show that gradient descent converges at a global linear rate to the global optimum. We believe these insights are also useful in analyzing deep models and other first order methods.

## Bayesian Policy Optimization for Model Uncertainty

- Gilwoo Lee, Brian Hou, Aditya Mandalika, Jeongseok Lee, Sanjiban Choudhury, Siddhartha S. Srinivasa
- abstract@open-review(Poster): Addressing uncertainty is critical for autonomous systems to robustly adapt to the real world. We formulate the problem of model uncertainty as a continuous Bayes-Adaptive Markov Decision Process (BAMDP), where an agent maintains a posterior distribution over latent model parameters given a history of observations and maximizes its expected long-term reward with respect to this belief distribution. Our algorithm, Bayesian Policy Optimization, builds on recent policy optimization algorithms to learn a universal policy that navigates the exploration-exploitation trade-off to maximize the Bayesian value function. To address challenges from discretizing the continuous latent parameter space, we propose a new policy network architecture that encodes the belief distribution independently from the observable state. Our method significantly outperforms algorithms that address model uncertainty without explicitly reasoning about belief distributions and is competitive with state-of-the-art Partially Observable Markov Decision Process solvers.

## Von Mises-Fisher Loss for Training Sequence to Sequence Models with Continuous Outputs

- Sachin Kumar, Yulia Tsvetkov
- abstract@open-review(Poster): The Softmax function is used in the final layer of nearly all existing sequence-to-sequence models for language generation. However, it is usually the slowest layer to compute which limits the vocabulary size to a subset of most frequent types; and it has a large memory footprint. We propose a general technique for replacing the softmax layer with a continuous embedding layer. Our primary innovations are a novel probabilistic loss, and a training and inference procedure in which we generate a probability distribution over pre-trained word embeddings, instead of a multinomial distribution over the vocabulary obtained via softmax. We evaluate this new class of sequence-to-sequence models with continuous outputs on the task of neural machine translation. We show that our models obtain upto 2.5x speed-up in training time while performing on par with the state-of-the-art models in terms of translation quality. These models are capable of handling very large vocabularies without compromising on translation quality. They also produce more meaningful errors than in the softmax-based models, as these errors typically lie in a subspace of the vector space of the reference translations.

## Information-Directed Exploration for Deep Reinforcement Learning

- Nikolay Nikolov, Johannes Kirschner, Felix Berkenkamp, Andreas Krause
- abstract@open-review(Poster): Efficient exploration remains a major challenge for reinforcement learning. One reason is that the variability of the returns often depends on the current state and action, and is therefore heteroscedastic. Classical exploration strategies such as upper confidence bound algorithms and Thompson sampling fail to appropriately account for heteroscedasticity, even in the bandit setting. Motivated by recent findings that address this issue in bandits, we propose to use Information-Directed Sampling (IDS) for exploration in reinforcement learning. As our main contribution, we build on recent advances in distributional reinforcement learning and propose a novel, tractable approximation of IDS for deep Q-learning. The resulting exploration strategy explicitly accounts for both parametric uncertainty and heteroscedastic observation noise. We evaluate our method on Atari games and demonstrate a significant improvement over alternative approaches.

## Learning deep representations by mutual information estimation and maximization

- R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, Yoshua Bengio
- abstract@open-review(Oral): This work investigates unsupervised learning of representations by maximizing mutual information between an input and the output of a deep neural network encoder. Importantly, we show that structure matters: incorporating knowledge about locality in the input into the objective can significantly improve a representation's suitability for downstream tasks. We further control characteristics of the representation by matching to a prior distribution adversarially. Our method, which we call Deep InfoMax (DIM), outperforms a number of popular unsupervised learning methods and compares favorably with fully-supervised learning on several classification tasks in with some standard architectures. DIM opens new avenues for unsupervised learning of representations and is an important step towards flexible formulations of representation learning objectives for specific end-goals.

## Generalized Tensor Models for Recurrent Neural Networks

- Valentin Khrulkov, Oleksii Hrinchuk, Ivan Oseledets
- abstract@open-review(Poster): Recurrent Neural Networks (RNNs) are very successful at solving challenging problems with sequential data. However, this observed efficiency is not yet entirely explained by theory. It is known that a certain class of multiplicative RNNs enjoys the property of depth

efficiency --- a shallow network of exponentially large width is necessary to realize the same score function as computed by such an RNN. Such networks, however, are not very often applied to real life tasks. In this work, we attempt to reduce the gap between theory and practice by extending the theoretical analysis to RNNs which employ various nonlinearities, such as Rectified Linear Unit (ReLU), and show that they also benefit from properties of universality and depth efficiency. Our theoretical results are verified by a series of extensive computational experiments.

## Invariant and Equivariant Graph Networks

- Haggai Maron, Heli Ben-Hamu, Nadav Shamir, Yaron Lipman
- abstract@open-review(Poster): Invariant and equivariant networks have been successfully used for learning images, sets, point clouds, and graphs. A basic challenge in developing such networks is finding the maximal collection of invariant and equivariant \emph{linear} layers. Although this question is answered for the first three examples (for popular transformations, at-least), a full characterization of invariant and equivariant linear layers for graphs is not known.

In this paper we provide a characterization of all permutation invariant and equivariant linear layers for (hyper-)graph data, and show that their dimension, in case of edge-value graph data, is $2$ and $15$, respectively. More generally, for graph data defined on $k$-tuples of nodes, the dimension is the $k$-th and $2k$-th Bell numbers. Orthogonal bases for the layers are computed, including generalization to multi-graph data. The constant number of basis elements and their characteristics allow successfully applying the networks to different size graphs. From the theoretical point of view, our results generalize and unify recent advancement in equivariant deep learning. In particular, we show that our model is capable of approximating any message passing neural network.

Applying these new linear layers in a simple deep neural network framework is shown to achieve comparable results to state-of-the-art and to have better expressivity than previous invariant and equivariant bases.

## Wizard of Wikipedia: Knowledge-Powered Conversational Agents

- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, Jason Weston
- abstract@open-review(Poster): In open-domain dialogue intelligent agents should exhibit the use of knowledge, however there are few convincing demonstrations of this to date. The most popular sequence to sequence models typically “generate and hope” generic utterances that can be memorized in the weights of the model when mapping from input utterance(s) to output, rather than employing recalled knowledge as context. Use of knowledge has so far proved difficult, in part because of the lack of a supervised learning benchmark task which exhibits knowledgeable open dialogue with clear grounding. To that end we collect and release a large dataset with conversations directly grounded with knowledge retrieved from Wikipedia. We then design architectures capable of retrieving knowledge, reading and conditioning on it, and finally generating natural responses. Our best performing dialogue models are able to conduct knowledgeable discussions on open-domain topics as evaluated by automatic metrics and human evaluations, while our new benchmark allows for measuring further improvements in this important research direction.

## Residual Non-local Attention Networks for Image Restoration

- Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, Yun Fu
- abstract@open-review(Poster): In this paper, we propose a residual non-local attention network for high-quality image restoration. Without considering the uneven distribution of information in the corrupted images, previous methods are restricted by local convolutional operation and equal treatment of spatial- and channel-wise features. To address this issue, we design local and non-local attention blocks to extract features that capture the long-range dependencies between pixels and pay more attention to the challenging parts. Specifically, we design trunk branch and (non-)local mask branch in each (non-)local attention block. The trunk branch is used to extract hierarchical features. Local and non-local mask branches aim to adaptively rescale these hierarchical features with mixed attentions. The local mask branch concentrates on more local structures with convolutional operations, while non-local attention considers more about long-range dependencies in the whole feature map. Furthermore, we propose residual local and non-local attention learning to train the very deep network, which further enhance the representation ability of the network. Our proposed method can be generalized for various image restoration applications, such as image denoising, demosaicing, compression artifacts reduction, and super-resolution. Experiments demonstrate that our method obtains comparable or better results compared with recently leading methods quantitatively and visually.

## Kernel RNN Learning (KeRNL)

- Christopher Roth, Ingmar Kanitscheider, Ila Fiete
- abstract@open-review(Poster): We describe Kernel RNN Learning (KeRNL), a reduced-rank, temporal eligibility trace-based approximation to backpropagation through time (BPTT) for training recurrent neural networks (RNNs) that gives competitive performance to BPTT on long time-dependence tasks. The approximation replaces a rank-4 gradient learning tensor, which describes how past hidden unit activations affect the current state, by a simple reduced-rank product of a sensitivity weight and a temporal eligibility trace. In this structured approximation motivated by node perturbation, the sensitivity weights and eligibility kernel time scales are themselves learned by applying perturbations. The rule represents another step toward biologically plausible or neurally inspired ML, with lower complexity in terms of relaxed architectural requirements (no symmetric return weights), a smaller memory demand (no unfolding and storage of states over time), and a shorter feedback time.

## Integer Networks for Data Compression with Latent-Variable Models

- Johannes Ballé, Nick Johnston, David Minnen
- abstract@open-review(Poster): We consider the problem of using variational latent-variable models for data compression. For such models to produce a compressed binary sequence, which is the universal data representation in a digital world, the latent representation needs to be subjected to entropy coding. Range coding as an entropy coding technique is optimal, but it can fail catastrophically if the computation of the prior differs even slightly between the sending and the receiving side. Unfortunately, this is a common scenario when floating point math is used and the sender and receiver operate on different hardware or software platforms, as numerical round-off is often platform dependent. We propose using integer networks as a universal solution to this problem, and demonstrate that they enable reliable cross-platform encoding and decoding of images using variational models.

## Structured Adversarial Attack: Towards General Implementation and Better Interpretability

- Kaidi Xu, Sijia Liu, Pu Zhao, Pin-Yu Chen, Huan Zhang, Quanfu Fan, Deniz Erdogmus, Yanzhi Wang, Xue Lin
- abstract@open-review(Poster): When generating adversarial examples to attack deep neural networks (DNNs), Lp norm of the added perturbation is usually used to measure the similarity between original image and adversarial example. However, such adversarial attacks perturbing the raw input spaces may fail to capture structural information hidden in the input. This work develops a more general attack model, i.e., the structured attack (StrAttack), which explores group sparsity in adversarial perturbation by sliding a mask through images aiming for extracting key spatial structures. An ADMM (alternating direction method of multipliers)-based framework is proposed that can split the original problem into a sequence of analytically solvable subproblems and can be generalized to implement other attacking methods. Strong group sparsity is achieved in adversarial perturbations even with the same level of Lp-norm distortion ($p \in \{1,2,\infty\}$) as the state-of-the-art attacks. We demonstrate the effectiveness of StrAttack by extensive experimental results on MNIST, CIFAR-10 and ImageNet. We also show that StrAttack provides better interpretability (i.e., better correspondence with discriminative image regions) through adversarial saliency map (Paper-not et al., 2016b) and class activation map (Zhou et al., 2016).

## Neural network gradient-based learning of black-box function interfaces

- Alon Jacovi, Guy Hadash, Einat Kermany, Boaz Carmeli, Ofer Lavi, George Kour, Jonathan Berant
- abstract@open-review(Poster): Deep neural networks work well at approximating complicated functions when provided with data and trained by gradient descent methods. At the same time, there is a vast amount of existing functions that programmatically solve different tasks in a precise manner eliminating the need for training. In many cases, it is possible to decompose a task to a series of functions, of which for some we may prefer to use a neural network to learn the functionality, while for others the preferred method would be to use existing black-box functions. We propose a method for end-to-end training of a base neural network that integrates calls to existing black-box functions. We do so by approximating the black-box functionality with a differentiable neural network in a way that drives the base network to comply with the black-box function interface during the end-to-end optimization process. At inference time, we replace the differentiable estimator with its external black-box non-differentiable counterpart such that the base network output matches the input arguments of the black-box function. Using this ``Estimate and Replace'' paradigm, we train a neural network, end to end, to compute the input to black-box functionality while eliminating the need for intermediate labels. We show that by leveraging the existing precise black-box function during inference, the integrated model generalizes better than a fully differentiable model, and learns more efficiently compared to RL-based methods.

## RNNs implicitly implement tensor-product representations

- R. Thomas McCoy, Tal Linzen, Ewan Dunbar, Paul Smolensky
- abstract@open-review(Poster): Recurrent neural networks (RNNs) can learn continuous vector representations of symbolic structures such as sequences and sentences; these representations often exhibit linear regularities (analogies). Such regularities motivate our hypothesis that RNNs that show such regularities implicitly compile symbolic structures into tensor product representations (TPRs; Smolensky, 1990), which additively combine tensor products of vectors representing roles (e.g., sequence positions) and vectors representing fillers (e.g., particular words). To test this hypothesis, we introduce Tensor Product Decomposition Networks (TPDNs), which use TPRs to approximate existing vector representations. We demonstrate using synthetic data that TPDNs can successfully approximate linear and tree-based RNN autoencoder representations, suggesting that these representations exhibit interpretable compositional structure; we explore the settings that lead RNNs to induce such structure-sensitive representations. By contrast, further TPDN experiments show that the representations of four models trained to encode naturally-occurring sentences can be largely approximated with a bag of words, with only marginal improvements from more sophisticated structures. We conclude that TPDNs provide a powerful method for interpreting vector representations, and that standard RNNs can induce compositional sequence representations that are remarkably well approximated byTPRs; at the same time, existing training tasks for sentence representation learning may not be sufficient for inducing robust structural representations

## Self-Monitoring Navigation Agent via Auxiliary Progress Estimation

- Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, Caiming Xiong
- abstract@open-review(Poster): The Vision-and-Language Navigation (VLN) task entails an agent following navigational instruction in photo-realistic unknown environments. This challenging task demands that the agent be aware of which instruction was completed, which instruction is needed next, which way to go, and its navigation progress towards the goal. In this paper, we introduce a self-monitoring agent with two complementary components: (1) visual-textual co-grounding module to locate the instruction completed in the past, the instruction required for the next action, and the next moving direction from surrounding images and (2) progress monitor to ensure the grounded instruction correctly reflects the navigation progress. We test our self-monitoring agent on a standard benchmark and analyze our proposed approach through a series of ablation studies that elucidate the contributions of the primary components. Using our proposed method, we set the new state of the art by a significant margin (8% absolute increase in success rate on the unseen test set). Code is available at https://github.com/chihyaoma/selfmonitoring-agent.

## Policy Transfer with Strategy Optimization

- Wenhao Yu, C. Karen Liu, Greg Turk
- abstract@open-review(Poster): Computer simulation provides an automatic and safe way for training robotic control policies to achieve complex tasks such as locomotion. However, a policy trained in simulation usually does not transfer directly to the real hardware due to the differences between the two environments. Transfer learning using domain randomization is a promising approach, but it usually assumes that the target environment is close to the distribution of the training environments, thus relying heavily on accurate system identification. In this paper, we present a different approach that leverages domain randomization for transferring control policies to unknown environments. The key idea that, instead of learning a single policy in the simulation, we simultaneously learn a family of policies that exhibit different behaviors. When tested in the target environment, we directly search for the best policy in the family based on the task performance, without the need to identify the dynamic parameters. We evaluate our method on five simulated robotic control problems with different discrepancies in the training and testing environment and demonstrate that our method can overcome larger modeling errors compared to training a robust policy or an adaptive policy.

## DeepOBS: A Deep Learning Optimizer Benchmark Suite

- Frank Schneider, Lukas Balles, Philipp Hennig
- abstract@open-review(Poster): Because the choice and tuning of the optimizer affects the speed, and ultimately the performance of deep learning, there is significant past and recent research in this area. Yet, perhaps surprisingly, there is no generally agreed-upon protocol for the quantitative and reproducible evaluation of optimization strategies for deep learning. We suggest routines and benchmarks for stochastic optimization, with special focus on the unique aspects of deep learning, such as stochasticity, tunability and generalization. As the primary contribution, we present DeepOBS, a Python package of deep learning optimization benchmarks. The package addresses key challenges in the quantitative assessment of stochastic optimizers, and automates most steps of benchmarking. The library includes a wide and extensible set of ready-to-use realistic optimization problems, such as training Residual Networks for image classification on ImageNet or character-level language prediction models, as well as popular classics like MNIST and CIFAR-10. The package also provides realistic baseline results for the most popular optimizers on these test problems, ensuring a fair comparison to the competition when benchmarking new optimizers, and without having to run costly experiments. It comes with output back-ends that directly produce LaTeX code for inclusion in academic publications. It supports TensorFlow and is available open source.

## signSGD with Majority Vote is Communication Efficient and Fault Tolerant

- Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, Anima Anandkumar
- abstract@open-review(Poster): Training neural networks on large datasets can be accelerated by distributing the workload over a network of machines. As datasets grow ever larger, networks of hundreds or thousands of machines become economically viable. The time cost of communicating gradients limits the effectiveness of using such large machine counts, as may the increased chance of network faults. We explore a particularly simple algorithm for robust, communication-efficient learning---signSGD. Workers transmit only the sign of their gradient vector to a server, and the overall update is decided by a majority vote. This algorithm uses 32x less communication per iteration than full-precision, distributed SGD. Under natural conditions verified by experiment, we prove that signSGD converges in the large and mini-batch settings, establishing convergence for a parameter regime of Adam as a byproduct. Aggregating sign gradients by majority vote means that no individual worker has too much power. We prove that unlike SGD, majority vote is robust when up to 50% of workers behave adversarially. The class of adversaries we consider includes as special cases those that invert or randomise their gradient estimate. On the practical side, we built our distributed training system in Pytorch. Benchmarking against the state of the art collective communications library (NCCL), our framework---with the parameter server housed entirely on one machine---led to a 25% reduction in time for training resnet50 on Imagenet when using 15 AWS p3.2xlarge machines.

## MARGINALIZED AVERAGE ATTENTIONAL NETWORK FOR WEAKLY-SUPERVISED LEARNING

- Yuan Yuan, Yueming Lyu, Xi Shen, Ivor W. Tsang, Dit-Yan Yeung
- abstract@open-review(Poster): In weakly-supervised temporal action localization, previous works have failed to locate dense and integral regions for each entire action due to the overestimation of the most salient regions. To alleviate this issue, we propose a marginalized average attentional network (MAAN) to suppress the dominant response of the most salient regions in a principled manner. The MAAN employs a novel marginalized average aggregation (MAA) module and learns a set of latent discriminative probabilities in an end-to-end fashion. MAA samples multiple subsets from the video snippet features according to a set of latent discriminative probabilities and takes the expectation over all the averaged subset features. Theoretically, we prove that the MAA module with learned latent discriminative probabilities successfully reduces the difference in responses between the most salient regions and the others. Therefore, MAAN is able to generate better class activation sequences and identify dense and integral action regions in the videos. Moreover, we propose a fast algorithm to reduce the complexity of constructing MAA from $O(2^T)$ to $O(T^2)$. Extensive experiments on two large-scale video datasets show that our MAAN achieves a superior performance on weakly-supervised temporal action localization.

## Learning Robust Representations by Projecting Superficial Statistics Out

- Haohan Wang, Zexue He, Zachary C. Lipton, Eric P. Xing
- abstract@open-review(Oral): Despite impressive performance as evaluated on i.i.d. holdout data, deep neural networks depend heavily on superficial statistics of the training data and are liable to break under distribution shift. For example, subtle changes to the background or texture of an image can break a seemingly powerful classifier. Building on previous work on domain generalization, we hope to produce a classifier that will generalize to previously unseen domains, even when domain identifiers are not available during training. This setting is challenging because the model may extract many distribution-specific (superficial) signals together with distribution-agnostic (semantic) signals. To overcome this challenge, we incorporate the gray-level co-occurrence matrix (GLCM) to extract patterns that our prior knowledge suggests are superficial: they are sensitive to the texture but unable to capture the gestalt of an image. Then we introduce two techniques for improving our networks' out-of-sample performance. The first method is built on the reverse gradient method that pushes our model to learn representations from which the GLCM representation is not predictable. The second method is built on the independence introduced by projecting the model's representation onto the subspace orthogonal to GLCM representation's. We test our method on the battery of standard domain generalization data sets and, interestingly, achieve comparable or better performance as compared to other domain generalization methods that explicitly require samples from the target distribution for training.

## Stable Opponent Shaping in Differentiable Games

- Alistair Letcher, Jakob Foerster, David Balduzzi, Tim Rocktäschel, Shimon Whiteson
- abstract@open-review(Poster): A growing number of learning methods are actually differentiable games whose players optimise multiple, interdependent objectives in parallel – from GANs and intrinsic curiosity to multi-agent RL. Opponent shaping is a powerful approach to improve learning dynamics in these games, accounting for player influence on others' updates. Learning with Opponent-Learning Awareness (LOLA) is a recent algorithm that exploits this response and leads to cooperation in settings like the Iterated Prisoner's Dilemma. Although experimentally successful, we show that LOLA agents can exhibit 'arrogant' behaviour directly at odds with convergence. In fact, remarkably few algorithms have theoretical guarantees applying across all (n-player, non-convex) games. In this paper we present Stable Opponent Shaping (SOS), a new method that interpolates between LOLA and a stable variant named LookAhead. We prove that LookAhead converges locally to equilibria and avoids strict saddles in all differentiable games. SOS inherits these essential guarantees, while also shaping the learning of opponents and consistently either matching or outperforming LOLA experimentally.

## Bounce and Learn: Modeling Scene Dynamics with Real-World Bounces

- Senthil Purushwalkam, Abhinav Gupta, Danny Kaufman, Bryan Russell
- abstract@open-review(Poster): We introduce an approach to model surface properties governing bounces in everyday scenes. Our model learns end-to-end, starting from sensor inputs, to predict post-bounce trajectories and infer two underlying physical properties that govern bouncing - restitution and effective collision normals. Our model, Bounce and Learn, comprises two modules -- a Physics Inference Module (PIM) and a Visual Inference Module (VIM). VIM learns to infer physical parameters for locations in a scene given a single still image, while PIM learns to model physical interactions for the prediction task given physical parameters and observed pre-collision 3D trajectories. To achieve our results, we introduce the Bounce Dataset comprising 5K RGB-D videos of bouncing trajectories of a foam ball to probe surfaces of varying shapes and materials in everyday scenes including homes and offices. Our proposed model learns from our collected dataset of real-world bounces and is bootstrapped with additional information from simple physics simulations. We show on our newly collected dataset that our model out-performs baselines, including trajectory fitting with Newtonian physics, in predicting post-bounce trajectories and inferring physical properties of a scene.

## Understanding Composition of Word Embeddings via Tensor Decomposition

- Abraham Frandsen, Rong Ge
- abstract@open-review(Poster): Word embedding is a powerful tool in natural language processing. In this paper we consider the problem of word embedding composition --- given vector representations of two words, compute a vector for the entire phrase. We give a generative model that can capture specific syntactic relations between words. Under our model, we prove that the correlations between three words (measured by their PMI) form a tensor that has an approximate low rank Tucker decomposition. The result of the Tucker decomposition gives the word embeddings as well as a core tensor, which can be used to produce better compositions of the word embeddings. We also complement our theoretical results with experiments that verify our assumptions, and demonstrate the effectiveness of the new composition method.

## SNAS: stochastic neural architecture search

- Sirui Xie, Hehui Zheng, Chunxiao Liu, Liang Lin
- abstract@open-review(Poster): We propose Stochastic Neural Architecture Search (SNAS), an economical end-to-end solution to Neural Architecture Search (NAS) that trains neural operation parameters and architecture distribution parameters in same round of back-propagation, while maintaining the completeness and differentiability of the NAS pipeline. In this work, NAS is reformulated as an optimization problem on parameters of a joint distribution for the search space in a cell. To leverage the gradient information in generic differentiable loss for architecture search, a novel search gradient is proposed. We prove that this search gradient optimizes the same objective as reinforcement-learning-based NAS, but assigns credits to structural decisions more efficiently. This credit assignment is further augmented with locally decomposable reward to enforce a resource-efficient constraint. In experiments on CIFAR-10, SNAS takes less epochs to find a cell architecture with state-of-the-art accuracy than non-differentiable evolution-based and reinforcement-learning-based NAS, which is also transferable to ImageNet. It is also shown that child networks of SNAS can maintain the validation accuracy in searching, with which attention-based NAS requires parameter retraining to compete, exhibiting potentials to stride towards efficient NAS on big datasets.

## Latent Convolutional Models

- ShahRukh Athar, Evgeny Burnaev, Victor Lempitsky
- abstract@open-review(Poster): We present a new latent model of natural images that can be learned on large-scale datasets. The learning process provides a latent embedding for every image in the training dataset, as well as a deep convolutional network that maps the latent space to the image space. After

training, the new model provides a strong and universal image prior for a variety of image restoration tasks such as large-hole inpainting, superresolution, and colorization. To model high-resolution natural images, our approach uses latent spaces of very high dimensionality (one to two orders of magnitude higher than previous latent image models). To tackle this high dimensionality, we use latent spaces with a special manifold structure (convolutional manifolds) parameterized by a ConvNet of a certain architecture. In the experiments, we compare the learned latent models with latent models learned by autoencoders, advanced variants of generative adversarial networks, and a strong baseline system using simpler parameterization of the latent space. Our model outperforms the competing approaches over a range of restoration tasks.

## NADPEx: An on-policy temporally consistent exploration method for deep reinforcement learning

- Sirui Xie, Junning Huang, Lanxin Lei, Chunxiao Liu, Zheng Ma, Wei Zhang, Liang Lin
- abstract@open-review(Poster): Reinforcement learning agents need exploratory behaviors to escape from local optima. These behaviors may include both immediate dithering perturbation and temporally consistent exploration. To achieve these, a stochastic policy model that is inherently consistent through a period of time is in desire, especially for tasks with either sparse rewards or long term information. In this work, we introduce a novel on-policy temporally consistent exploration strategy - Neural Adaptive Dropout Policy Exploration (NADPEx) - for deep reinforcement learning agents. Modeled as a global random variable for conditional distribution, dropout is incorporated to reinforcement learning policies, equipping them with inherent temporal consistency, even when the reward signals are sparse. Two factors, gradients' alignment with the objective and KL constraint in policy space, are discussed to guarantee NADPEx policy's stable improvement. Our experiments demonstrate that NADPEx solves tasks with sparse reward while naive exploration and parameter noise fail. It yields as well or even faster convergence in the standard mujoco benchmark for continuous control.

## Representation Degeneration Problem in Training Natural Language Generation Models

- Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, Tieyan Liu
- abstract@open-review(Poster): We study an interesting problem in training neural network-based models for natural language generation tasks, which we call the \emph{representation degeneration problem}. We observe that when training a model for natural language generation tasks through likelihood maximization with the weight tying trick, especially with big training datasets, most of the learnt word embeddings tend to degenerate and be distributed into a narrow cone, which largely limits the representation power of word embeddings. We analyze the conditions and causes of this problem and propose a novel regularization method to address it. Experiments on language modeling and machine translation show that our method can largely mitigate the representation degeneration problem and achieve better performance than baseline algorithms.

## Soft Q-Learning with Mutual-Information Regularization

- Jordi Grau-Moya, Felix Leibfried, Peter Vrancx
- abstract@open-review(Poster): We propose a reinforcement learning (RL) algorithm that uses mutual-information regularization to optimize a prior action distribution for better performance and exploration. Entropy-based regularization has previously been shown to improve both exploration and robustness in challenging sequential decision-making tasks. It does so by encouraging policies to put probability mass on all actions. However, entropy regularization might be undesirable when actions have significantly different importance. In this paper, we propose a theoretically motivated framework that dynamically weights the importance of actions by using the mutual-information. In particular, we express the RL problem as an inference problem where the prior probability distribution over actions is subject to optimization. We show that the prior optimization introduces a mutual-information regularizer in the RL objective. This regularizer encourages the policy to be close to a non-uniform distribution that assigns higher probability mass to more important actions. We empirically demonstrate that our method significantly improves over entropy regularization methods and unregularized methods.

## Learning to Adapt in Dynamic, Real-World Environments through Meta-Reinforcement Learning

- Anusha Nagabandi, Ignasi Clavera, Simin Liu, Ronald S. Fearing, Pieter Abbeel, Sergey Levine, Chelsea Finn
- abstract@open-review(Poster): Although reinforcement learning methods can achieve impressive results in simulation, the real world presents two major challenges: generating samples is exceedingly expensive, and unexpected perturbations or unseen situations cause proficient but specialized policies to fail at test time. Given that it is impractical to train separate policies to accommodate all situations the agent may see in the real world, this work proposes to learn how to quickly and effectively adapt online to new tasks. To enable sample-efficient learning, we consider learning online adaptation in the context of model-based reinforcement learning. Our approach uses meta-learning to train a dynamics model prior such that, when combined with recent data, this prior can be rapidly adapted to the local context. Our experiments demonstrate online adaptation for continuous control tasks on both simulated and real-world agents. We first show simulated agents adapting their behavior online to novel terrains, crippled body parts, and highly-dynamic environments. We also illustrate the importance of incorporating online adaptation into autonomous agents that operate in the real world by applying our method to a real dynamic legged millirobot: We demonstrate the agent's learned ability to quickly adapt online to a missing leg, adjust to novel terrains and slopes, account for miscalibration or errors in pose estimation, and compensate for pulling payloads.

## How Important is a Neuron

- Kedar Dhamdhere, Mukund Sundararajan, Qiqi Yan
- abstract@open-review(Poster): The problem of attributing a deep network’s prediction to its input/base features is well-studied (cf. Simonyan et al. (2013)). We introduce the notion of conductance to extend the notion of attribution to understanding the importance of hidden units. Informally, the conductance of a hidden unit of a deep network is the flow of attribution via this hidden unit. We can use conductance to understand the importance of a hidden unit to the prediction for a specific input, or over a set of inputs. We justify conductance in multiple ways via a qualitative comparison with other methods, via some axiomatic results, and via an empirical evaluation based on a feature selection task. The empirical evaluations are done using the Inception network over ImageNet data, and a convolutinal network over text data. In both cases, we demonstrate the effectiveness of conductance in identifying interesting insights about the internal workings of these networks.

## Knowledge Flow: Improve Upon Your Teachers

- Iou-Jen Liu, Jian Peng, Alexander Schwing
- abstract@open-review(Poster): A zoo of deep nets is available these days for almost any given task, and it is increasingly unclear which net to start with when addressing a new task, or which net to use as an initialization for fine-tuning a new model. To address this issue, in this paper, we develop knowledge flow which moves ‘knowledge’ from multiple deep nets, referred to as teachers, to a new deep net model, called the student. The structure of the teachers and the student can differ arbitrarily and they can be trained on entirely different tasks with different output spaces too. Upon training with knowledge flow the student is independent of the teachers. We demonstrate our approach on a variety of supervised and reinforcement learning tasks, outperforming fine-tuning and other ‘knowledge exchange’ methods.

## Meta-Learning Update Rules for Unsupervised Representation Learning

- Luke Metz, Niru Maheswaranathan, Brian Cheung, Jascha Sohl-Dickstein
- abstract@open-review(Oral): A major goal of unsupervised learning is to discover data representations that are useful for subsequent tasks, without access to supervised labels during training. Typically, this involves minimizing a surrogate objective, such as the negative log likelihood of a generative model,

with the hope that representations useful for subsequent tasks will arise as a side effect. In this work, we propose instead to directly target later desired tasks by meta-learning an unsupervised learning rule which leads to representations useful for those tasks. Specifically, we target semi-supervised classification performance, and we meta-learn an algorithm -- an unsupervised weight update rule -- that produces representations useful for this task. Additionally, we constrain our unsupervised update rule to a be a biologically-motivated, neuron-local function, which enables it to generalize to different neural network architectures, datasets, and data modalities. We show that the meta-learned update rule produces useful features and sometimes outperforms existing unsupervised learning techniques. We further show that the meta-learned unsupervised update rule generalizes to train networks with different widths, depths, and nonlinearities. It also generalizes to train on data with randomly permuted input dimensions and even generalizes from image datasets to a text task.

## Emergent Coordination Through Competition

- Siqi Liu, Guy Lever, Josh Merel, Saran Tunyasuvunakool, Nicolas Heess, Thore Graepel
- abstract@open-review(Poster): We study the emergence of cooperative behaviors in reinforcement learning agents by introducing a challenging competitive multi-agent soccer environment with continuous simulated physics. We demonstrate that decentralized, population-based training with co-play can lead to a progression in agents' behaviors: from random, to simple ball chasing, and finally showing evidence of cooperation. Our study highlights several of the challenges encountered in large scale multi-agent training in continuous control. In particular, we demonstrate that the automatic optimization of simple shaping rewards, not themselves conducive to co-operative behavior, can lead to long-horizon team behavior. We further apply an evaluation scheme, grounded by game theoretic principals, that can assess agent performance in the absence of pre-defined evaluation tasks or human baselines.

## Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile

- Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, Georgios Piliouras
- abstract@open-review(Poster): Owing to their connection with generative adversarial networks (GANs), saddle-point problems have recently attracted considerable interest in machine learning and beyond. By necessity, most theoretical guarantees revolve around convex-concave (or even linear) problems; however, making theoretical inroads towards efficient GAN training depends crucially on moving beyond this classic framework. To make piecemeal progress along these lines, we analyze the behavior of mirror descent (MD) in a class of non-monotone problems whose solutions coincide with those of a naturally associated variational inequality – a property which we call coherence. We first show that ordinary, “vanilla” MD converges under a strict version of this condition, but not otherwise; in particular, it may fail to converge even in bilinear models with a unique solution. We then show that this deficiency is mitigated by optimism: by taking an “extra-gradient” step, optimistic mirror descent (OMD) converges in all coherent problems. Our analysis generalizes and extends the results of Daskalakis et al. [2018] for optimistic gradient descent (OGD) in bilinear problems, and makes concrete headway for provable convergence beyond convex-concave games. We also provide stochastic analogues of these results, and we validate our analysis by numerical experiments in a wide array of GAN models (including Gaussian mixture models, and the CelebA and CIFAR-10 datasets).

## Multilingual Neural Machine Translation with Knowledge Distillation

- Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, Tie-Yan Liu
- abstract@open-review(Poster): Multilingual machine translation, which translates multiple languages with a single model, has attracted much attention due to its efficiency of offline training and online serving. However, traditional multilingual translation usually yields inferior accuracy compared with the counterpart using individual models for each language pair, due to language diversity and model capacity limitations. In this paper, we propose a distillation-based approach to boost the accuracy of multilingual machine translation. Specifically, individual models are first trained and regarded as teachers, and then the multilingual model is trained to fit the training data and match the outputs of individual models simultaneously through knowledge distillation. Experiments on IWSLT, WMT and Ted talk translation datasets demonstrate the effectiveness of our method. Particularly, we show that one model is enough to handle multiple languages (up to 44 languages in our experiment), with comparable or even better accuracy than individual models.

## Structured Neural Summarization

- Patrick Fernandes, Miltiadis Allamanis, Marc Brockschmidt
- abstract@open-review(Poster): Summarization of long sequences into a concise statement is a core problem in natural language processing, requiring non-trivial understanding of the input. Based on the promising results of graph neural networks on highly structured data, we develop a framework to extend existing sequence encoders with a graph component that can reason about long-distance relationships in weakly structured data such as text. In an extensive evaluation, we show that the resulting hybrid sequence-graph models outperform both pure sequence models as well as pure graph models on a range of summarization tasks.

## Hyperbolic Attention Networks

- Caglar Gulcehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter Battaglia, Victor Bapst, David Raposo, Adam Santoro, Nando de Freitas
- abstract@open-review(Poster): Recent approaches have successfully demonstrated the benefits of learning the parameters of shallow networks in hyperbolic space. We extend this line of work by imposing hyperbolic geometry on the embeddings used to compute the ubiquitous attention mechanisms for different neural networks architectures. By only changing the geometry of embedding of object representations, we can use the embedding space more efficiently without increasing the number of parameters of the model. Mainly as the number of objects grows exponentially for any semantic distance from the query, hyperbolic geometry --as opposed to Euclidean geometry-- can encode those objects without having any interference. Our method shows improvements in generalization on neural machine translation on WMT'14 (English to German), learning on graphs (both on synthetic and real-world graph tasks) and visual question answering (CLEVR) tasks while keeping the neural representations compact.

## Deep Learning 3D Shapes Using Alt-az Anisotropic 2-Sphere Convolution

- Min Liu, Fupin Yao, Chiho Choi, Ayan Sinha, Karthik Ramani
- abstract@open-review(Poster): The ground-breaking performance obtained by deep convolutional neural networks (CNNs) for image processing tasks is inspiring research efforts attempting to extend it for 3D geometric tasks. One of the main challenge in applying CNNs to 3D shape analysis is how to define a natural convolution operator on non-euclidean surfaces. In this paper, we present a method for applying deep learning to 3D surfaces using their spherical descriptors and alt-az anisotropic convolution on 2-sphere. A cascade set of geodesic disk filters rotate on the 2-sphere and collect spherical patterns and so to extract geometric features for various 3D shape analysis tasks. We demonstrate theoretically and experimentally that our proposed method has the possibility to bridge the gap between 2D images and 3D shapes with the desired rotation equivariance/invariance, and its effectiveness is evaluated in applications of non-rigid/ rigid shape classification and shape retrieval.

## Generative predecessor models for sample-efficient imitation learning

- Yannick Schroecker, Mel Vecerik, Jon Scholz

- abstract@open-review(Poster): We propose Generative Predecessor Models for Imitation Learning (GPRIL), a novel imitation learning algorithm that matches the state-action distribution to the distribution observed in expert demonstrations, using generative models to reason probabilistically about alternative histories of demonstrated states. We show that this approach allows an agent to learn robust policies using only a small number of expert demonstrations and self-supervised interactions with the environment. We derive this approach from first principles and compare it empirically to a state-of-the-art imitation learning method, showing that it outperforms or matches its performance on two simulated robot manipulation tasks and demonstrate significantly higher sample efficiency by applying the algorithm on a real robot.

## Relational Forward Models for Multi-Agent Learning

- Andrea Tacchetti, H. Francis Song, Pedro A. M. Mediano, Vinicius Zambaldi, János Kramár, Neil C. Rabinowitz, Thore Graepel, Matthew Botvinick, Peter W. Battaglia
- abstract@open-review(Poster): The behavioral dynamics of multi-agent systems have a rich and orderly structure, which can be leveraged to understand these systems, and to improve how artificial agents learn to operate in them. Here we introduce Relational Forward Models (RFM) for multi-agent learning, networks that can learn to make accurate predictions of agents' future behavior in multi-agent environments. Because these models operate on the discrete entities and relations present in the environment, they produce interpretable intermediate representations which offer insights into what drives agents' behavior, and what events mediate the intensity and valence of social interactions. Furthermore, we show that embedding RFM modules inside agents results in faster learning systems compared to non-augmented baselines. As more and more of the autonomous systems we develop and interact with become multi-agent in nature, developing richer analysis tools for characterizing how and why agents make decisions is increasingly necessary. Moreover, developing artificial agents that quickly and safely learn to coordinate with one another, and with humans in shared environments, is crucial.

## Variational Bayesian Phylogenetic Inference

- Cheng Zhang, Frederick A. Matsen IV
- abstract@open-review(Poster): Bayesian phylogenetic inference is currently done via Markov chain Monte Carlo with simple mechanisms for proposing new states, which hinders exploration efficiency and often requires long runs to deliver accurate posterior estimates. In this paper we present an alternative approach: a variational framework for Bayesian phylogenetic analysis. We approximate the true posterior using an expressive graphical model for tree distributions, called a subsplit Bayesian network, together with appropriate branch length distributions. We train the variational approximation via stochastic gradient ascent and adopt multi-sample based gradient estimators for different latent variables separately to handle the composite latent space of phylogenetic models. We show that our structured variational approximations are flexible enough to provide comparable posterior estimation to MCMC, while requiring less computation due to a more efficient tree exploration mechanism enabled by variational inference. Moreover, the variational approximations can be readily used for further statistical analysis such as marginal likelihood estimation for model comparison via importance sampling. Experiments on both synthetic data and real data Bayesian phylogenetic inference problems demonstrate the effectiveness and efficiency of our methods.

## Adv-BNN: Improved Adversarial Defense through Robust Bayesian Neural Network

- Xuanqing Liu, Yao Li, Chongruo Wu, Cho-Jui Hsieh
- abstract@open-review(Poster): We present a new algorithm to train a robust neural network against adversarial attacks. Our algorithm is motivated by the following two ideas. First, although recent work has demonstrated that fusing randomness can improve the robustness of neural networks (Liu 2017), we noticed that adding noise blindly to all the layers is not the optimal way to incorporate randomness. Instead, we model randomness under the framework of Bayesian Neural Network (BNN) to formally learn the posterior distribution of models in a scalable way. Second, we formulate the mini-max problem in BNN to learn the best model distribution under adversarial attacks, leading to an adversarial-trained Bayesian neural net. Experiment results demonstrate that the proposed algorithm achieves state-of-the-art performance under strong attacks. On CIFAR-10 with VGG network, our model leads to 14% accuracy improvement compared with adversarial training (Madry 2017) and random self-ensemble (Liu, 2017) under PGD attack with 0.035 distortion, and the gap becomes even larger on a subset of ImageNet.

## h-detach: Modifying the LSTM Gradient Towards Better Optimization

- Bhargav Kanuparthi, Devansh Arpit, Giancarlo Kerg, Nan Rosemary Ke, Ioannis Mitliagkas, Yoshua Bengio
- abstract@open-review(Poster): Recurrent neural networks are known for their notorious exploding and vanishing gradient problem (EVGP). This problem becomes more evident in tasks where the information needed to correctly solve them exist over long time scales, because EVGP prevents important gradient components from being back-propagated adequately over a large number of steps. We introduce a simple stochastic algorithm (\textit{h}-detach) that is specific to LSTM optimization and targeted towards addressing this problem. Specifically, we show that when the LSTM weights are large, the gradient components through the linear path (cell state) in the LSTM computational graph get suppressed. Based on the hypothesis that these components carry information about long term dependencies (which we show empirically), their suppression can prevent LSTMs from capturing them. Our algorithm\footnote{Our code is available at https://github.com/bhargav104/h-detach.} prevents gradients flowing through this path from getting suppressed, thus allowing the LSTM to capture such dependencies better. We show significant improvements over vanilla LSTM gradient based training in terms of convergence speed, robustness to seed and learning rate, and generalization using our modification of LSTM gradient on various benchmark datasets.

## On the loss landscape of a class of deep neural networks with no bad local valleys

- Quynh Nguyen, Mahesh Chandra Mukkamala, Matthias Hein
- abstract@open-review(Poster): We identify a class of over-parameterized deep neural networks with standard activation functions and cross-entropy loss which provably have no bad local valley, in the sense that from any point in parameter space there exists a continuous path on which the cross-entropy loss is non-increasing and gets arbitrarily close to zero. This implies that these networks have no sub-optimal strict local minima.

## Accumulation Bit-Width Scaling For Ultra-Low Precision Training Of Deep Networks

- Charbel Sakr, Naigang Wang, Chia-Yu Chen, Jungwook Choi, Ankur Agrawal, Naresh Shanbhag, Kailash Gopalakrishnan
- abstract@open-review(Poster): Efforts to reduce the numerical precision of computations in deep learning training have yielded systems that aggressively quantize weights and activations, yet employ wide high-precision accumulators for partial sums in inner-product operations to preserve the quality of convergence. The absence of any framework to analyze the precision requirements of partial sum accumulations results in conservative design choices. This imposes an upper-bound on the reduction of complexity of multiply-accumulate units. We present a statistical approach to analyze the impact of reduced accumulation precision on deep learning training. Observing that a bad choice for accumulation precision results in loss of information that manifests itself as a reduction in variance in an ensemble of partial sums, we derive a set of equations that relate this variance to the length of accumulation and the minimum number of bits needed for accumulation. We apply our analysis to three benchmark networks: CIFAR-10 ResNet 32, ImageNet ResNet 18 and ImageNet AlexNet. In each case, with accumulation precision set in accordance with our proposed equations, the networks successfully converge to the single precision floating-point baseline. We also show that reducing accumulation precision further degrades the quality of the trained network, proving that our equations produce tight bounds. Overall this analysis enables precise tailoring of computation hardware to the application, yielding area- and power-optimal systems.

## Deep Convolutional Networks as shallow Gaussian Processes

- Adrià Garriga-Alonso, Carl Edward Rasmussen, Laurence Aitchison
- abstract@open-review(Poster): We show that the output of a (residual) CNN with an appropriate prior over the weights and biases is a GP in the limit of infinitely many convolutional filters, extending similar results for dense networks. For a CNN, the equivalent kernel can be computed exactly and, unlike "deep kernels", has very few parameters: only the hyperparameters of the original CNN. Further, we show that this kernel has two properties that allow it to be computed efficiently; the cost of evaluating the kernel for a pair of images is similar to a single forward pass through the original CNN with only one filter per layer. The kernel equivalent to a 32-layer ResNet obtains 0.84% classification error on MNIST, a new record for GP with a comparable number of parameters.

## Universal Successor Features Approximators

- Diana Borsa, Andre Barreto, John Quan, Daniel J. Mankowitz, Hado van Hasselt, Remi Munos, David Silver, Tom Schaul
- abstract@open-review(Poster): The ability of a reinforcement learning (RL) agent to learn about many reward functions at the same time has many potential benefits, such as the decomposition of complex tasks into simpler ones, the exchange of information between tasks, and the reuse of skills. We focus on one aspect in particular, namely the ability to generalise to unseen tasks. Parametric generalisation relies on the interpolation power of a function approximator that is given the task description as input; one of its most common form are universal value function approximators (UVFAs). Another way to generalise to new tasks is to exploit structure in the RL problem itself. Generalised policy improvement (GPI) combines solutions of previous tasks into a policy for the unseen task; this relies on instantaneous policy evaluation of old policies under the new reward function, which is made possible through successor features (SFs). Our proposed \emph{universal successor features approximators} (USFAs) combine the advantages of all of these, namely the scalability of UVFAs, the instant inference of SFs, and the strong generalisation of GPI. We discuss the challenges involved in training a USFA, its generalisation properties and demonstrate its practical benefits and transfer abilities on a large-scale domain in which the agent has to navigate in a first-person perspective three-dimensional environment.

## Adaptive Estimators Show Information Compression in Deep Neural Networks

- Ivan Chelombiev, Conor Houghton, Cian O'Donnell
- abstract@open-review(Poster): To improve how neural networks function it is crucial to understand their learning process. The information bottleneck theory of deep learning proposes that neural networks achieve good generalization by compressing their representations to disregard information that is not relevant to the task. However, empirical evidence for this theory is conflicting, as compression was only observed when networks used saturating activation functions. In contrast, networks with non-saturating activation functions achieved comparable levels of task performance but did not show compression. In this paper we developed more robust mutual information estimation techniques, that adapt to hidden activity of neural networks and produce more sensitive measurements of activations from all functions, especially unbounded functions. Using these adaptive estimation techniques, we explored compression in networks with a range of different activation functions. With two improved methods of estimation, firstly, we show that saturation of the activation function is not required for compression, and the amount of compression varies between different activation functions. We also find that there is a large amount of variation in compression between different network initializations. Secondary, we see that L2 regularization leads to significantly increased compression, while preventing overfitting. Finally, we show that only compression of the last layer is positively correlated with generalization.

## CBOW Is Not All You Need: Combining CBOW with the Compositional Matrix Space Model

- Florian Mai, Lukas Galke, Ansgar Scherp
- abstract@open-review(Poster): Continuous Bag of Words (CBOW) is a powerful text embedding method. Due to its strong capabilities to encode word content, CBOW embeddings perform well on a wide range of downstream tasks while being efficient to compute. However, CBOW is not capable of capturing the word order. The reason is that the computation of CBOW's word embeddings is commutative, i.e., embeddings of XYZ and ZYX are the same. In order to address this shortcoming, we propose a learning algorithm for the Continuous Matrix Space Model, which we call Continual Multiplication of Words (CMOW). Our algorithm is an adaptation of word2vec, so that it can be trained on large quantities of unlabeled text. We empirically show that CMOW better captures linguistic properties, but it is inferior to CBOW in memorizing word content. Motivated by these findings, we propose a hybrid model that combines the strengths of CBOW and CMOW. Our results show that the hybrid CBOW-CMOW-model retains CBOW's strong ability to memorize word content while at the same time substantially improving its ability to encode other linguistic information by 8%. As a result, the hybrid also performs better on 8 out of 11 supervised downstream tasks with an average improvement of 1.2%.

## Fluctuation-dissipation relations for stochastic gradient descent

- Sho Yaida
- abstract@open-review(Poster): The notion of the stationary equilibrium ensemble has played a central role in statistical mechanics. In machine learning as well, training serves as generalized equilibration that drives the probability distribution of model parameters toward stationarity. Here, we derive stationary fluctuation-dissipation relations that link measurable quantities and hyperparameters in the stochastic gradient descent algorithm. These relations hold exactly for any stationary state and can in particular be used to adaptively set training schedule. We can further use the relations to efficiently extract information pertaining to a loss-function landscape such as the magnitudes of its Hessian and anharmonicity. Our claims are empirically verified.

## A Universal Music Translation Network

- Noam Mor, Lior Wolf, Adam Polyak, Yaniv Taigman
- abstract@open-review(Poster): We present a method for translating music across musical instruments and styles. This method is based on unsupervised training of a multi-domain wavenet autoencoder, with a shared encoder and a domain-independent latent space that is trained end-to-end on waveforms. Employing a diverse training dataset and large net capacity, the single encoder allows us to translate also from musical domains that were not seen during training. We evaluate our method on a dataset collected from professional musicians, and achieve convincing translations. We also study the properties of the obtained translation and demonstrate translating even from a whistle, potentially enabling the creation of instrumental music by untrained humans.

## Neural Persistence: A Complexity Measure for Deep Neural Networks Using Algebraic Topology

- Bastian Rieck, Matteo Togninalli, Christian Bock, Michael Moor, Max Horn, Thomas Gumbsch, Karsten Borgwardt
- abstract@open-review(Poster): While many approaches to make neural networks more fathomable have been proposed, they are restricted to interrogating the network with input data. Measures for characterizing and monitoring structural properties, however, have not been developed. In this work, we propose neural persistence, a complexity measure for neural network architectures based on topological data analysis on weighted stratified graphs. To demonstrate the usefulness of our approach, we show that neural persistence reflects best practices developed in the deep learning community such as dropout and batch normalization. Moreover, we derive a neural persistence-based stopping criterion that shortens the training process while achieving comparable accuracies as early stopping based on validation loss.

## Detecting Egregious Responses in Neural Sequence-to-sequence Models

- Tianxing He, James Glass
- abstract@open-review(Poster): In this work, we attempt to answer a critical question: whether there exists some input sequence that will cause a well-trained discrete-space neural network sequence-to-sequence (seq2seq) model to generate egregious outputs (aggressive, malicious, attacking, etc.). And if such inputs exist, how to find them efficiently. We adopt an empirical methodology, in which we first create lists of egregious output sequences, and then design a discrete optimization algorithm to find input sequences that will cause the model to generate them. Moreover, the optimization algorithm is enhanced for large vocabulary search and constrained to search for input sequences that are likely to be input by real-world users. In our experiments, we apply this approach to dialogue response generation models trained on three real-world dialogue data-sets: Ubuntu, Switchboard and OpenSubtitles, testing whether the model can generate malicious responses. We demonstrate that given the trigger inputs our algorithm finds, a significant number of malicious sentences are assigned large probability by the model, which reveals an undesirable consequence of standard seq2seq training.

## Measuring Compositionality in Representation Learning

- Jacob Andreas
- abstract@open-review(Poster): Many machine learning algorithms represent input data with vector embeddings or discrete codes. When inputs exhibit compositional structure (e.g. objects built from parts or procedures from subroutines), it is natural to ask whether this compositional structure is reflected in the the inputs' learned representations. While the assessment of compositionality in languages has received significant attention in linguistics and adjacent fields, the machine learning literature lacks general-purpose tools for producing graded measurements of compositional structure in more general (e.g. vector-valued) representation spaces. We describe a procedure for evaluating compositionality by measuring how well the true representation-producing model can be approximated by a model that explicitly composes a collection of inferred representational primitives. We use the procedure to provide formal and empirical characterizations of compositional structure in a variety of settings, exploring the relationship between compositionality and learning dynamics, human judgments, representational similarity, and generalization.

## Learning Representations of Sets through Optimized Permutations

- Yan Zhang, Jonathon Hare, Adam Prügel-Bennett
- abstract@open-review(Poster): Representations of sets are challenging to learn because operations on sets should be permutation-invariant. To this end, we propose a Permutation-Optimisation module that learns how to permute a set end-to-end. The permuted set can be further processed to learn a permutation-invariant representation of that set, avoiding a bottleneck in traditional set models. We demonstrate our model's ability to learn permutations and set representations with either explicit or implicit supervision on four datasets, on which we achieve state-of-the-art results: number sorting, image mosaics, classification from image mosaics, and visual question answering.

## Analysing Mathematical Reasoning Abilities of Neural Models

- David Saxton, Edward Grefenstette, Felix Hill, Pushmeet Kohli
- abstract@open-review(Poster): Mathematical reasoning---a core ability within human intelligence---presents some unique challenges as a domain: we do not come to understand and solve mathematical problems primarily on the back of experience and evidence, but on the basis of inferring, learning, and exploiting laws, axioms, and symbol manipulation rules. In this paper, we present a new challenge for the evaluation (and eventually the design) of neural architectures and similar system, developing a task suite of mathematics problems involving sequential questions and answers in a free-form textual input/output format. The structured nature of the mathematics domain, covering arithmetic, algebra, probability and calculus, enables the construction of training and test spits designed to clearly illuminate the capabilities and failure-modes of different architectures, as well as evaluate their ability to compose and relate knowledge and learned processes. Having described the data generation process and its potential future expansions, we conduct a comprehensive analysis of models from two broad classes of the most powerful sequence-to-sequence architectures and find notable differences in their ability to resolve mathematical problems and generalize their knowledge.

## Ordered Neurons: Integrating Tree Structures into Recurrent Neural Networks

- Yikang Shen, Shawn Tan, Alessandro Sordoni, Aaron Courville
- abstract@open-review(Oral): Natural language is hierarchically structured: smaller units (e.g., phrases) are nested within larger units (e.g., clauses). When a larger constituent ends, all of the smaller constituents that are nested within it must also be closed. While the standard LSTM architecture allows different neurons to track information at different time scales, it does not have an explicit bias towards modeling a hierarchy of constituents. This paper proposes to add such inductive bias by ordering the neurons; a vector of master input and forget gates ensures that when a given neuron is updated, all the neurons that follow it in the ordering are also updated. Our novel recurrent architecture, ordered neurons LSTM (ON-LSTM), achieves good performance on four different tasks: language modeling, unsupervised parsing, targeted syntactic evaluation, and logical inference.

## FUNCTIONAL VARIATIONAL BAYESIAN NEURAL NETWORKS

- Shengyang Sun, Guodong Zhang, Jiaxin Shi, Roger Grosse
- abstract@open-review(Poster): Variational Bayesian neural networks (BNN) perform variational inference over weights, but it is difficult to specify meaningful priors and approximating posteriors in a high-dimensional weight space. We introduce functional variational Bayesian neural networks (fBNNs), which maximize an Evidence Lower BOund (ELBO) defined directly on stochastic processes, i.e. distributions over functions. We prove that the KL divergence between stochastic processes is equal to the supremum of marginal KL divergences over all finite sets of inputs. Based on this, we introduce a practical training objective which approximates the functional ELBO using finite measurement sets and the spectral Stein gradient estimator. With fBNNs, we can specify priors which entail rich structure, including Gaussian processes and implicit stochastic processes. Empirically, we find that fBNNs extrapolate well using various structured priors, provide reliable uncertainty estimates, and can scale to large datasets.

## ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness

- Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, Wieland Brendel
- abstract@open-review(Oral): Convolutional Neural Networks (CNNs) are commonly thought to recognise objects by learning increasingly complex representations of object shapes. Some recent studies suggest a more important role of image textures. We here put these conflicting hypotheses to a quantitative test by evaluating CNNs and human observers on images with a texture-shape cue conflict. We show that ImageNet-trained CNNs are strongly biased towards recognising textures rather than shapes, which is in stark contrast to human behavioural evidence and reveals fundamentally different classification strategies. We then demonstrate that the same standard architecture (ResNet-50) that learns a texture-based representation on ImageNet is able to learn a shape-based representation instead when trained on 'Stylized-ImageNet', a stylized version of ImageNet. This provides a much better fit for human behavioural performance in our well-controlled psychophysical lab setting (nine experiments totalling 48,560 psychophysical trials across 97 observers) and comes with a number of unexpected emergent benefits such as improved object detection performance and previously unseen robustness towards a wide range of image distortions, highlighting advantages of a shape-based representation.

## Improving MMD-GAN Training with Repulsive Loss Function

- Wei Wang, Yuan Sun, Saman Halgamuge
- abstract@open-review(Poster): Generative adversarial nets (GANs) are widely used to learn the data sampling process and their performance may heavily depend on the loss functions, given a limited computational budget. This study revisits MMD-GAN that uses the maximum mean discrepancy (MMD) as the loss function for GAN and makes two contributions. First, we argue that the existing MMD loss function may discourage the learning of fine details in data as it attempts to contract the discriminator outputs of real data. To address this issue, we propose a repulsive loss function to actively learn the difference among the real data by simply rearranging the terms in MMD. Second, inspired by the hinge loss, we propose a bounded Gaussian kernel to stabilize the training of MMD-GAN with the repulsive loss function. The proposed methods are applied to the unsupervised image generation tasks on CIFAR-10, STL-10, CelebA, and LSUN bedroom datasets. Results show that the repulsive loss function significantly improves over the MMD loss at no additional computational cost and outperforms other representative loss functions. The proposed methods achieve an FID score of 16.21 on the CIFAR-10 dataset using a single DCGAN network and spectral normalization.

## Large Scale GAN Training for High Fidelity Natural Image Synthesis

- Andrew Brock, Jeff Donahue, Karen Simonyan
- abstract@open-review(Oral): Despite recent progress in generative image modeling, successfully generating high-resolution, diverse samples from complex datasets such as ImageNet remains an elusive goal. To this end, we train Generative Adversarial Networks at the largest scale yet attempted, and study the instabilities specific to such scale. We find that applying orthogonal regularization to the generator renders it amenable to a simple "truncation trick", allowing fine control over the trade-off between sample fidelity and variety by reducing the variance of the Generator's input. Our modifications lead to models which set the new state of the art in class-conditional image synthesis. When trained on ImageNet at 128x128 resolution, our models (BigGANs) achieve an Inception Score (IS) of 166.3 and Frechet Inception Distance (FID) of 9.6, improving over the previous best IS of 52.52 and FID of 18.65.

## SOM-VAE: Interpretable Discrete Representation Learning on Time Series

- Vincent Fortuin, Matthias Hüser, Francesco Locatello, Heiko Strathmann, Gunnar Rätsch
- abstract@open-review(Poster): High-dimensional time series are common in many domains. Since human cognition is not optimized to work well in high-dimensional spaces, these areas could benefit from interpretable low-dimensional representations. However, most representation learning algorithms for time series data are difficult to interpret. This is due to non-intuitive mappings from data features to salient properties of the representation and non-smoothness over time. To address this problem, we propose a new representation learning framework building on ideas from interpretable discrete dimensionality reduction and deep generative modeling. This framework allows us to learn discrete representations of time series, which give rise to smooth and interpretable embeddings with superior clustering performance. We introduce a new way to overcome the non-differentiability in discrete representation learning and present a gradient-based version of the traditional self-organizing map algorithm that is more performant than the original. Furthermore, to allow for a probabilistic interpretation of our method, we integrate a Markov model in the representation space. This model uncovers the temporal transition structure, improves clustering performance even further and provides additional explanatory insights as well as a natural representation of uncertainty. We evaluate our model in terms of clustering performance and interpretability on static (Fashion-)MNIST data, a time series of linearly interpolated (Fashion-)MNIST images, a chaotic Lorenz attractor system with two macro states, as well as on a challenging real world medical time series application on the eICU data set. Our learned representations compare favorably with competitor methods and facilitate downstream tasks on the real world data.

## Riemannian Adaptive Optimization Methods

- Gary Becigneul, Octavian-Eugen Ganea
- abstract@open-review(Poster): Several first order stochastic optimization methods commonly used in the Euclidean domain such as stochastic gradient descent (SGD), accelerated gradient descent or variance reduced methods have already been adapted to certain Riemannian settings. However, some of the most popular of these optimization tools - namely Adam, Adagrad and the more recent Amsgrad - remain to be generalized to Riemannian manifolds. We discuss the difficulty of generalizing such adaptive schemes to the most agnostic Riemannian setting, and then provide algorithms and convergence proofs for geodesically convex objectives in the particular case of a product of Riemannian manifolds, in which adaptivity is implemented across manifolds in the cartesian product. Our generalization is tight in the sense that choosing the Euclidean space as Riemannian manifold yields the same algorithms and regret bounds as those that were already known for the standard algorithms. Experimentally, we show faster convergence and to a lower train loss value for Riemannian adaptive methods over their corresponding baselines on the realistic task of embedding the WordNet taxonomy in the Poincare ball.

## Non-vacuous Generalization Bounds at the ImageNet Scale: a PAC-Bayesian Compression Approach

- Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, Peter Orbanz
- abstract@open-review(Poster): Modern neural networks are highly overparameterized, with capacity to substantially overfit to training data. Nevertheless, these networks often generalize well in practice. It has also been observed that trained networks can often be ``compressed to much smaller representations. The purpose of this paper is to connect these two empirical observations. Our main technical result is a generalization bound for compressed networks based on the compressed size that, combined with off-the-shelf compression algorithms, leads to state-of-the-art generalization guarantees. In particular, we provide the first non-vacuous generalization guarantees for realistic architectures applied to the ImageNet classification problem. Additionally, we show that compressibility of models that tend to overfit is limited. Empirical results show that an increase in overfitting increases the number of bits required to describe a trained network.

## Learning Factorized Multimodal Representations

- Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, Ruslan Salakhutdinov
- abstract@open-review(Poster): Learning multimodal representations is a fundamentally complex research problem due to the presence of multiple heterogeneous sources of information. Although the presence of multiple modalities provides additional valuable information, there are two key challenges to address when learning from multimodal data: 1) models must learn the complex intra-modal and cross-modal interactions for prediction and 2) models must be robust to unexpected missing or noisy modalities during testing. In this paper, we propose to optimize for a joint generative-discriminative objective across multimodal data and labels. We introduce a model that factorizes representations into two sets of independent factors: multimodal discriminative and modality-specific generative factors. Multimodal discriminative factors are shared across all modalities and contain joint multimodal features required for discriminative tasks such as sentiment prediction. Modality-specific generative factors are unique for each modality and contain the information required for generating data. Experimental results show that our model is able to learn meaningful multimodal representations that achieve state-of-the-art or competitive performance on six multimodal datasets. Our model demonstrates flexible generative capabilities by conditioning on independent factors and can reconstruct missing modalities without significantly impacting performance. Lastly, we interpret our factorized representations to understand the interactions that influence multimodal learning.

## Aggregated Momentum: Stability Through Passive Damping

- James Lucas, Shengyang Sun, Richard Zemel, Roger Grosse
- abstract@open-review(Poster): Momentum is a simple and widely used trick which allows gradient-based optimizers to pick up speed along low curvature directions. Its performance depends crucially on a damping coefficient. Largecamping coefficients can potentially deliver much larger speedups, but are

prone to oscillations and instability; hence one typically resorts to small values such as 0.5 or 0.9. We propose Aggregated Momentum (AggMo), a variant of momentum which combines multiple velocity vectors with different damping coefficients. AggMo is trivial to implement, but significantly dampens oscillations, enabling it to remain stable even for aggressive damping coefficients such as 0.999. We reinterpret Nesterov's accelerated gradient descent as a special case of AggMo and analyze rates of convergence for quadratic objectives. Empirically, we find that AggMo is a suitable drop-in replacement for other momentum methods, and frequently delivers faster convergence with little to no tuning.

## Learning to Schedule Communication in Multi-agent Reinforcement Learning

- Daewoo Kim, Sangwoo Moon, David Hostallero, Wan Ju Kang, Taeyoung Lee, Kyunghwan Son, Yung Yi
- abstract@open-review(Poster): Many real-world reinforcement learning tasks require multiple agents to make sequential decisions under the agents' interaction, where well-coordinated actions among the agents are crucial to achieve the target goal better at these tasks. One way to accelerate the coordination effect is to enable multiple agents to communicate with each other in a distributed manner and behave as a group. In this paper, we study a practical scenario when (i) the communication bandwidth is limited and (ii) the agents share the communication medium so that only a restricted number of agents are able to simultaneously use the medium, as in the state-of-the-art wireless networking standards. This calls for a certain form of communication scheduling. In that regard, we propose a multi-agent deep reinforcement learning framework, called SchedNet, in which agents learn how to schedule themselves, how to encode the messages, and how to select actions based on received messages. SchedNet is capable of deciding which agents should be entitled to broadcasting their (encoded) messages, by learning the importance of each agent's partially observed information. We evaluate SchedNet against multiple baselines under two different applications, namely, cooperative communication and navigation, and predator-prey. Our experiments show a non-negligible performance gap between SchedNet and other mechanisms such as the ones without communication and with vanilla scheduling methods, e.g., round robin, ranging from 32% to 43%.

## Minimum Divergence vs. Maximum Margin: an Empirical Comparison on Seq2Seq Models

- Huan Zhang, Hai Zhao
- abstract@open-review(Poster): Sequence to sequence (seq2seq) models have become a popular framework for neural sequence prediction. While traditional seq2seq models are trained by Maximum Likelihood Estimation (MLE), much recent work has made various attempts to optimize evaluation scores directly to solve the mismatch between training and evaluation, since model predictions are usually evaluated by a task specific evaluation metric like BLEU or ROUGE scores instead of perplexity. This paper puts this existing work into two categories, a) minimum divergence, and b) maximum margin. We introduce a new training criterion based on the analysis of existing work, and empirically compare models in the two categories. Our experimental results show that our new training criterion can usually work better than existing methods, on both the tasks of machine translation and sentence summarization.

## Overcoming Catastrophic Forgetting for Continual Learning via Model Adaptation

- Wenpeng Hu, Zhou Lin, Bing Liu, Chongyang Tao, Zhengwei Tao, Jinwen Ma, Dongyan Zhao, Rui Yan
- abstract@open-review(Poster): Learning multiple tasks sequentially is important for the development of AI and lifelong learning systems. However, standard neural network architectures suffer from catastrophic forgetting which makes it difficult for them to learn a sequence of tasks. Several continual learning methods have been proposed to address the problem. In this paper, we propose a very different approach, called Parameter Generation and Model Adaptation (PGMA), to dealing with the problem. The proposed approach learns to build a model, called the solver, with two sets of parameters. The first set is shared by all tasks learned so far and the second set is dynamically generated to adapt the solver to suit each test example in order to classify it. Extensive experiments have been carried out to demonstrate the effectiveness of the proposed approach.

## Multi-Domain Adversarial Learning

- Alice Schoenauer-Sebag, Louise Heinrich, Marc Schoenauer, Michele Sebag, Lani F. Wu, Steve J. Altschuler
- abstract@open-review(Poster): Multi-domain learning (MDL) aims at obtaining a model with minimal average risk across multiple domains. Our empirical motivation is automated microscopy data, where cultured cells are imaged after being exposed to known and unknown chemical perturbations, and each dataset displays significant experimental bias. This paper presents a multi-domain adversarial learning approach, MuLANN, to leverage multiple datasets with overlapping but distinct class sets, in a semi-supervised setting. Our contributions include: i) a bound on the average- and worst-domain risk in MDL, obtained using the H-divergence; ii) a new loss to accommodate semi-supervised multi-domain learning and domain adaptation; iii) the experimental validation of the approach, improving on the state of the art on two standard image benchmarks, and a novel bioimage dataset, Cell.

## Learning from Positive and Unlabeled Data with a Selection Bias

- Masahiro Kato, Takeshi Teshima, Junya Honda
- abstract@open-review(Poster): We consider the problem of learning a binary classifier only from positive data and unlabeled data (PU learning). Recent methods of PU learning commonly assume that the labeled positive data are identically distributed as the unlabeled positive data. However, this assumption is unrealistic in many instances of PU learning because it fails to capture the existence of a selection bias in the labeling process. When the data has a selection bias, it is difficult to learn the Bayes optimal classifier by conventional methods of PU learning. In this paper, we propose a method to partially identify the classifier. The proposed algorithm learns a scoring function that preserves the order induced by the class posterior under mild assumptions, which can be used as a classifier by setting an appropriate threshold. Through experiments, we show that the method outperforms previous methods for PU learning on various real-world datasets.

## CEM-RL: Combining evolutionary and gradient-based methods for policy search

- Pourchot, Sigaud
- abstract@open-review(Poster): Deep neuroevolution and deep reinforcement learning (deep RL) algorithms are two popular approaches to policy search. The former is widely applicable and rather stable, but suffers from low sample efficiency. By contrast, the latter is more sample efficient, but the most sample efficient variants are also rather unstable and highly sensitive to hyper-parameter setting. So far, these families of methods have mostly been compared as competing tools. However, an emerging approach consists in combining them so as to get the best of both worlds. Two previously existing combinations use either an ad hoc evolutionary algorithm or a goal exploration process together with the Deep Deterministic Policy Gradient (DDPG) algorithm, a sample efficient off-policy deep RL algorithm. In this paper, we propose a different combination scheme using the simple cross-entropy method (CEM) and Twin Delayed Deep Deterministic policy gradient (TD3), another off-policy deep RL algorithm which improves over DDPG. We evaluate the resulting method, CEM-RL, on a set of benchmarks classically used in deep RL. We show that CEM-RL benefits from several advantages over its competitors and offers a satisfactory trade-off between performance and sample efficiency.

## SGD Converges to Global Minimum in Deep Learning via Star-convex Path

- Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, Vahid Tarokh
- abstract@open-review(Poster): Stochastic gradient descent (SGD) has been found to be surprisingly effective in training a variety of deep neural networks. However, there is still a lack of understanding on how and why SGD can train these complex networks towards a global minimum. In this study,

we establish the convergence of SGD to a global minimum for nonconvex optimization problems that are commonly encountered in neural network training. Our argument exploits the following two important properties: 1) the training loss can achieve zero value (approximately), which has been widely observed in deep learning; 2) SGD follows a star-convex path, which is verified by various experiments in this paper. In such a context, our analysis shows that SGD, although has long been considered as a randomized algorithm, converges in an intrinsically deterministic manner to a global minimum.

## LayoutGAN: Generating Graphic Layouts with Wireframe Discriminators

- Jianan Li, Jimei Yang, Aaron Hertzmann, Jianming Zhang, Tingfa Xu
- abstract@open-review(Poster): Layout is important for graphic design and scene generation. We propose a novel Generative Adversarial Network, called LayoutGAN, that synthesizes layouts by modeling geometric relations of different types of 2D elements. The generator of LayoutGAN takes as input a set of randomly-placed 2D graphic elements and uses self-attention modules to refine their labels and geometric parameters jointly to produce a realistic layout. Accurate alignment is critical for good layouts. We thus propose a novel differentiable wireframe rendering layer that maps the generated layout to a wireframe image, upon which a CNN-based discriminator is used to optimize the layouts in image space. We validate the effectiveness of LayoutGAN in various experiments including MNIST digit generation, document layout generation, clipart abstract scene generation and tangram graphic design.

## Equi-normalization of Neural Networks

- Pierre Stock, Benjamin Graham, Rémi Gribonval, Hervé Jégou
- abstract@open-review(Poster): Modern neural networks are over-parametrized. In particular, each rectified linear hidden unit can be modified by a multiplicative factor by adjusting input and out- put weights, without changing the rest of the network. Inspired by the Sinkhorn-Knopp algorithm, we introduce a fast iterative method for minimizing the l2 norm of the weights, equivalently the weight decay regularizer. It provably converges to a unique solution. Interleaving our algorithm with SGD during training improves the test accuracy. For small batches, our approach offers an alternative to batch- and group- normalization on CIFAR-10 and ImageNet with a ResNet-18.

## Information Theoretic lower bounds on negative log likelihood

- Luis A. Lastras-Montaño
- abstract@open-review(Poster): In this article we use rate-distortion theory, a branch of information theory devoted to the problem of lossy compression, to shed light on an important problem in latent variable modeling of data: is there room to improve the model? One way to address this question is to find an upper bound on the probability (equivalently a lower bound on the negative log likelihood) that the model can assign to some data as one varies the prior and/or the likelihood function in a latent variable model. The core of our contribution is to formally show that the problem of optimizing priors in latent variable models is exactly an instance of the variational optimization problem that information theorists solve when computing rate-distortion functions, and then to use this to derive a lower bound on negative log likelihood. Moreover, we will show that if changing the prior can improve the log likelihood, then there is a way to change the likelihood function instead and attain the same log likelihood, and thus rate-distortion theory is of relevance to both optimizing priors as well as optimizing likelihood functions. We will experimentally argue for the usefulness of quantities derived from rate-distortion theory in latent variable modeling by applying them to a problem in image modeling.

## Neural Speed Reading with Structural-Jump-LSTM

- Christian Hansen, Casper Hansen, Stephen Alstrup, Jakob Grue Simonsen, Christina Lioma
- abstract@open-review(Poster): Recurrent neural networks (RNNs) can model natural language by sequentially "reading" input tokens and outputting a distributed representation of each token. Due to the sequential nature of RNNs, inference time is linearly dependent on the input length, and all inputs are read regardless of their importance. Efforts to speed up this inference, known as "neural speed reading", either ignore or skim over part of the input. We present Structural-Jump-LSTM: the first neural speed reading model to both skip and jump text during inference. The model consists of a standard LSTM and two agents: one capable of skipping single words when reading, and one capable of exploiting punctuation structure (sub-sentence separators (,:), sentence end symbols (.!?), or end of text markers) to jump ahead after reading a word. A comprehensive experimental evaluation of our model against all five state-of-the-art neural reading models shows that Structural-Jump-LSTM achieves the best overall floating point operations (FLOP) reduction (hence is faster), while keeping the same accuracy or even improving it compared to a vanilla LSTM that reads the whole text.

## Deep Graph Infomax

- Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, R Devon Hjelm
- abstract@open-review(Poster): We present Deep Graph Infomax (DGI), a general approach for learning node representations within graph-structured data in an unsupervised manner. DGI relies on maximizing mutual information between patch representations and corresponding high-level summaries of graphs---both derived using established graph convolutional network architectures. The learnt patch representations summarize subgraphs centered around nodes of interest, and can thus be reused for downstream node-wise learning tasks. In contrast to most prior approaches to unsupervised learning with GCNs, DGI does not rely on random walk objectives, and is readily applicable to both transductive and inductive learning setups. We demonstrate competitive performance on a variety of node classification benchmarks, which at times even exceeds the performance of supervised learning.

## SNIP: SINGLE-SHOT NETWORK PRUNING BASED ON CONNECTION SENSITIVITY

- Namhoon Lee, Thalaiyasingam Ajanthan, Philip Torr
- abstract@open-review(Poster): Pruning large neural networks while maintaining their performance is often desirable due to the reduced space and time complexity. In existing methods, pruning is done within an iterative optimization procedure with either heuristically designed pruning schedules or additional hyperparameters, undermining their utility. In this work, we present a new approach that prunes a given network once at initialization prior to training. To achieve this, we introduce a saliency criterion based on connection sensitivity that identifies structurally important connections in the network for the given task. This eliminates the need for both pretraining and the complex pruning schedule while making it robust to architecture variations. After pruning, the sparse network is trained in the standard way. Our method obtains extremely sparse networks with virtually the same accuracy as the reference network on the MNIST, CIFAR-10, and Tiny-ImageNet classification tasks and is broadly applicable to various architectures including convolutional, residual and recurrent networks. Unlike existing methods, our approach enables us to demonstrate that the retained connections are indeed relevant to the given task.

## Bias-Reduced Uncertainty Estimation for Deep Neural Classifiers

- Yonatan Geifman, Guy Uziel, Ran El-Yaniv
- abstract@open-review(Poster): We consider the problem of uncertainty estimation in the context of (non-Bayesian) deep neural classification. In this context, all known methods are based on extracting uncertainty signals from a trained network optimized to solve the classification problem at hand. We demonstrate that such techniques tend to introduce biased estimates for instances whose predictions are supposed to be highly confident. We argue that this deficiency is an artifact of the dynamics of training with SGD-like optimizers, and it has some properties similar to overfitting. Based on this observation, we develop an uncertainty estimation algorithm that selectively estimates the uncertainty of highly confident points, using earlier snapshots of

the trained model, before their estimates are jittered (and way before they are ready for actual classification). We present extensive experiments indicating that the proposed algorithm provides uncertainty estimates that are consistently better than all known methods.

## Approximability of Discriminators Implies Diversity in GANs

- Yu Bai, Tengyu Ma, Andrej Risteski
- abstract@open-review(Poster): While Generative Adversarial Networks (GANs) have empirically produced impressive results on learning complex real-world distributions, recent works have shown that they suffer from lack of diversity or mode collapse. The theoretical work of Arora et al. (2017a) suggests a dilemma about GANs' statistical properties: powerful discriminators cause overfitting, whereas weak discriminators cannot detect mode collapse. By contrast, we show in this paper that GANs can in principle learn distributions in Wasserstein distance (or KL-divergence in many cases) with polynomial sample complexity, if the discriminator class has strong distinguishing power against the particular generator class (instead of against all possible generators). For various generator classes such as mixture of Gaussians, exponential families, and invertible and injective neural networks generators, we design corresponding discriminators (which are often neural nets of specific architectures) such that the Integral Probability Metric (IPM) induced by the discriminators can provably approximate the Wasserstein distance and/or KL-divergence. This implies that if the training is successful, then the learned distribution is close to the true distribution in Wasserstein distance or KL divergence, and thus cannot drop modes. Our preliminary experiments show that on synthetic datasets the test IPM is well correlated with KL divergence or the Wasserstein distance, indicating that the lack of diversity in GANs may be caused by the sub-optimality in optimization instead of statistical inefficiency.

## On the Minimal Supervision for Training Any Binary Classifier from Only Unlabeled Data

- Nan Lu, Gang Niu, Aditya Krishna Menon, Masashi Sugiyama
- abstract@open-review(Poster): Empirical risk minimization (ERM), with proper loss function and regularization, is the common practice of supervised classification. In this paper, we study training arbitrary (from linear to deep) binary classifier from only unlabeled (U) data by ERM. We prove that it is impossible to estimate the risk of an arbitrary binary classifier in an unbiased manner given a single set of U data, but it becomes possible given two sets of U data with different class priors. These two facts answer a fundamental question---what the minimal supervision is for training any binary classifier from only U data. Following these findings, we propose an ERM-based learning method from two sets of U data, and then prove it is consistent. Experiments demonstrate the proposed method could train deep models and outperform state-of-the-art methods for learning from two sets of U data.

## KnockoffGAN: Generating Knockoffs for Feature Selection using Generative Adversarial Networks

- James Jordon, Jinsung Yoon, Mihaela van der Schaar
- abstract@open-review(Oral): Feature selection is a pervasive problem. The discovery of relevant features can be as important for performing a particular task (such as to avoid overfitting in prediction) as it can be for understanding the underlying processes governing the true label (such as discovering relevant genetic factors for a disease). Machine learning driven feature selection can enable discovery from large, high-dimensional, non-linear observational datasets by creating a subset of features for experts to focus on. In order to use expert time most efficiently, we need a principled methodology capable of controlling the False Discovery Rate. In this work, we build on the promising Knockoff framework by developing a flexible knockoff generation model. We adapt the Generative Adversarial Networks framework to allow us to generate knockoffs with no assumptions on the feature distribution. Our model consists of 4 networks, a generator, a discriminator, a stability network and a power network. We demonstrate the capability of our model to perform feature selection, showing that it performs as well as the originally proposed knockoff generation model in the Gaussian setting and that it outperforms the original model in non-Gaussian settings, including on a real-world dataset.

## Auxiliary Variational MCMC

- Raza Habib, David Barber
- abstract@open-review(Poster): We introduce Auxiliary Variational MCMC, a novel framework for learning MCMC kernels that combines recent advances in variational inference with insights drawn from traditional auxiliary variable MCMC methods such as Hamiltonian Monte Carlo. Our framework exploits low dimensional structure in the target distribution in order to learn a more efficient MCMC sampler. The resulting sampler is able to suppress random walk behaviour and mix between modes efficiently, without the need to compute gradients of the target distribution. We test our sampler on a number of challenging distributions, where the underlying structure is known, and on the task of posterior sampling in Bayesian logistic regression. Code to reproduce all experiments is available at https://github.com/AVMCMC/AuxiliaryVariationalMCMC .

## PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees

- James Jordon, Jinsung Yoon, Mihaela van der Schaar
- abstract@open-review(Poster): Machine learning has the potential to assist many communities in using the large datasets that are becoming more and more available. Unfortunately, much of that potential is not being realized because it would require sharing data in a way that compromises privacy. In this paper, we investigate a method for ensuring (differential) privacy of the generator of the Generative Adversarial Nets (GAN) framework. The resulting model can be used for generating synthetic data on which algorithms can be trained and validated, and on which competitions can be conducted, without compromising the privacy of the original dataset. Our method modifies the Private Aggregation of Teacher Ensembles (PATE) framework and applies it to GANs. Our modified framework (which we call PATE-GAN) allows us to tightly bound the influence of any individual sample on the model, resulting in tight differential privacy guarantees and thus an improved performance over models with the same guarantees. We also look at measuring the quality of synthetic data from a new angle; we assert that for the synthetic data to be useful for machine learning researchers, the relative performance of two algorithms (trained and tested) on the synthetic dataset should be the same as their relative performance (when trained and tested) on the original dataset. Our experiments, on various datasets, demonstrate that PATE-GAN consistently outperforms the state-of-the-art method with respect to this and other notions of synthetic data quality.

## Minimal Random Code Learning: Getting Bits Back from Compressed Model Parameters

- Marton Havasi, Robert Peharz, José Miguel Hernández-Lobato
- abstract@open-review(Poster): While deep neural networks are a highly successful model class, their large memory footprint puts considerable strain on energy consumption, communication bandwidth, and storage requirements. Consequently, model size reduction has become an utmost goal in deep learning. A typical approach is to train a set of deterministic weights, while applying certain techniques such as pruning and quantization, in order that the empirical weight distribution becomes amenable to Shannon-style coding schemes. However, as shown in this paper, relaxing weight determinism and using a full variational distribution over weights allows for more efficient coding schemes and consequently higher compression rates. In particular, following the classical bits-back argument, we encode the network weights using a random sample, requiring only a number of bits corresponding to the Kullback-Leibler divergence between the sampled variational distribution and the encoding distribution. By imposing a constraint on the Kullback-Leibler divergence, we are able to explicitly control the compression rate, while optimizing the expected loss on the training set. The employed encoding scheme can be shown to be close to the optimal information-theoretical lower bound, with respect to the employed variational family. Our method sets new state-of-the-art in neural network compression, as it strictly dominates previous approaches in a Pareto sense: On the benchmarks LeNet-5/MNIST and VGG-16/CIFAR-10, our approach yields the best test performance for a fixed memory budget, and vice versa, it achieves the highest compression rates for a fixed test performance.

## GO Gradient for Expectation-Based Objectives

- Yulai Cong, Miaoyun Zhao, Ke Bai, Lawrence Carin
- abstract@open-review(Poster): Within many machine learning algorithms, a fundamental problem concerns efficient calculation of an unbiased gradient wrt parameters $\boldsymbol{\gamma}$ for expectation-based objectives $\mathbb{E}_{q(\boldsymbol{y})} [f (\boldsymbol{y}) ]$. Most existing methods either ($i$) suffer from high variance, seeking help from (often) complicated variance-reduction techniques; or ($ii$) they only apply to reparameterizable continuous random variables and employ a reparameterization trick. To address these limitations, we propose a General and One-sample (GO) gradient that ($i$) applies to many distributions associated with non-reparameterizable continuous {\em or} discrete random variables, and ($ii$) has the same low-variance as the reparameterization trick. We find that the GO gradient often works well in practice based on only one Monte Carlo sample (although one can of course use more samples if desired). Alongside the GO gradient, we develop a means of propagating the chain rule through distributions, yielding statistical back-propagation, coupling neural networks to common random variables.

## Benchmarking Neural Network Robustness to Common Corruptions and Perturbations

- Dan Hendrycks, Thomas Dietterich
- abstract@open-review(Poster): In this paper we establish rigorous benchmarks for image classifier robustness. Our first benchmark, ImageNet-C, standardizes and expands the corruption robustness topic, while showing which classifiers are preferable in safety-critical applications. Then we propose a new dataset called ImageNet-P which enables researchers to benchmark a classifier's robustness to common perturbations. Unlike recent robustness research, this benchmark evaluates performance on common corruptions and perturbations not worst-case adversarial perturbations. We find that there are negligible changes in relative corruption robustness from AlexNet classifiers to ResNet classifiers. Afterward we discover ways to enhance corruption and perturbation robustness. We even find that a bypassed adversarial defense provides substantial common perturbation robustness. Together our benchmarks may aid future work toward networks that robustly generalize.

## Deep reinforcement learning with relational inductive biases

- Vinicius Zambaldi, David Raposo, Adam Santoro, Victor Bapst, Yujia Li, Igor Babuschkin, Karl Tuyls, David Reichert, Timothy Lillicrap, Edward Lockhart, Murray Shanahan, Victoria Langston, Razvan Pascanu, Matthew Botvinick, Oriol Vinyals, Peter Battaglia
- abstract@open-review(Poster): We introduce an approach for augmenting model-free deep reinforcement learning agents with a mechanism for relational reasoning over structured representations, which improves performance, learning efficiency, generalization, and interpretability. Our architecture encodes an image as a set of vectors, and applies an iterative message-passing procedure to discover and reason about relevant entities and relations in a scene. In six of seven StarCraft II Learning Environment mini-games, our agent achieved state-of-the-art performance, and surpassed human grandmaster-level on four. In a novel navigation and planning task, our agent's performance and learning efficiency far exceeded non-relational baselines, it was able to generalize to more complex scenes than it had experienced during training. Moreover, when we examined its learned internal representations, they reflected important structure about the problem and the agent's intentions. The main contribution of this work is to introduce techniques for representing and reasoning about states in model-free deep reinforcement learning agents via relational inductive biases. Our experiments show this approach can offer advantages in efficiency, generalization, and interpretability, and can scale up to meet some of the most challenging test environments in modern artificial intelligence.

## L-Shapley and C-Shapley: Efficient Model Interpretation for Structured Data

- Jianbo Chen, Le Song, Martin J. Wainwright, Michael I. Jordan
- abstract@open-review(Poster): Instancewise feature scoring is a method for model interpretation, which yields, for each test instance, a vector of importance scores associated with features. Methods based on the Shapley score have been proposed as a fair way of computing feature attributions, but incur an exponential complexity in the number of features. This combinatorial explosion arises from the definition of Shapley value and prevents these methods from being scalable to large data sets and complex models. We focus on settings in which the data have a graph structure, and the contribution of features to the target variable is well-approximated by a graph-structured factorization. In such settings, we develop two algorithms with linear complexity for instancewise feature importance scoring on black-box models. We establish the relationship of our methods to the Shapley value and a closely related concept known as the Myerson value from cooperative game theory. We demonstrate on both language and image data that our algorithms compare favorably with other methods using both quantitative metrics and human evaluation.

## Whitening and Coloring Batch Transform for GANs

- Aliaksandr Siarohin, Enver Sangineto, Nicu Sebe
- abstract@open-review(Poster): Batch Normalization (BN) is a common technique used to speed-up and stabilize training. On the other hand, the learnable parameters of BN are commonly used in conditional Generative Adversarial Networks (cGANs) for representing class-specific information using conditional Batch Normalization (cBN). In this paper we propose to generalize both BN and cBN using a Whitening and Coloring based batch normalization. We show that our conditional Coloring can represent categorical conditioning information which largely helps the cGAN qualitative results. Moreover, we show that full-feature whitening is important in a general GAN scenario in which the training process is known to be highly unstable. We test our approach on different datasets and using different GAN networks and training protocols, showing a consistent improvement in all the tested frameworks. Our CIFAR-10 conditioned results are higher than all previous works on this dataset.

## PeerNets: Exploiting Peer Wisdom Against Adversarial Attacks

- Jan Svoboda, Jonathan Masci, Federico Monti, Michael Bronstein, Leonidas Guibas
- abstract@open-review(Poster): Deep learning systems have become ubiquitous in many aspects of our lives. Unfortunately, it has been shown that such systems are vulnerable to adversarial attacks, making them prone to potential unlawful uses. Designing deep neural networks that are robust to adversarial attacks is a fundamental step in making such systems safer and deployable in a broader variety of applications (e.g. autonomous driving), but more importantly is a necessary step to design novel and more advanced architectures built on new computational paradigms rather than marginally building on the existing ones. In this paper we introduce PeerNets, a novel family of convolutional networks alternating classical Euclidean convolutions with graph convolutions to harness information from a graph of peer samples. This results in a form of non-local forward propagation in the model, where latent features are conditioned on the global structure induced by the graph, that is up to 3 times more robust to a variety of white- and black-box adversarial attacks compared to conventional architectures with almost no drop in accuracy.

## Sparse Dictionary Learning by Dynamical Neural Networks

- Tsung-Han Lin, Ping Tak Peter Tang
- abstract@open-review(Poster): A dynamical neural network consists of a set of interconnected neurons that interact over time continuously. It can exhibit computational properties in the sense that the dynamical system’s evolution and/or limit points in the associated state space can correspond to numerical solutions to certain mathematical optimization or learning problems. Such a computational system is particularly attractive in that it can be mapped to a massively parallel computer architecture for power and throughput efficiency, especially if each neuron can rely solely on local information (i.e., local memory). Deriving gradients from the dynamical network’s various states while conforming to this last constraint, however, is challenging. We show that

by combining ideas of top-down feedback and contrastive learning, a dynamical network for solving the l1-minimizing dictionary learning problem can be constructed, and the true gradients for learning are provably computable by individual neurons. Using spiking neurons to construct our dynamical network, we present a learning process, its rigorous mathematical analysis, and numerical results on several dictionary learning problems.

## Relaxed Quantization for Discretized Neural Networks

- Christos Louizos, Matthias Reisser, Tijmen Blankevoort, Efstratios Gavves, Max Welling
- abstract@open-review(Poster): Neural network quantization has become an important research area due to its great impact on deployment of large models on resource constrained devices. In order to train networks that can be effectively discretized without loss of performance, we introduce a differentiable quantization procedure. Differentiability can be achieved by transforming continuous distributions over the weights and activations of the network to categorical distributions over the quantization grid. These are subsequently relaxed to continuous surrogates that can allow for efficient gradient-based optimization. We further show that stochastic rounding can be seen as a special case of the proposed approach and that under this formulation the quantization grid itself can also be optimized with gradient descent. We experimentally validate the performance of our method on MNIST, CIFAR 10 and Imagenet classification.

## Marginal Policy Gradients: A Unified Family of Estimators for Bounded Action Spaces with Applications

- Carson Eisenach, Haichuan Yang, Ji Liu, Han Liu
- abstract@open-review(Poster): Many complex domains, such as robotics control and real-time strategy (RTS) games, require an agent to learn a continuous control. In the former, an agent learns a policy over R^d and in the latter, over a discrete set of actions each of which is parametrized by a continuous parameter. Such problems are naturally solved using policy based reinforcement learning (RL) methods, but unfortunately these often suffer from high variance leading to instability and slow convergence. Unnecessary variance is introduced whenever policies over bounded action spaces are modeled using distributions with unbounded support by applying a transformation T to the sampled action before execution in the environment. Recently, the variance reduced clipped action policy gradient (CAPG) was introduced for actions in bounded intervals, but to date no variance reduced methods exist when the action is a direction, something often seen in RTS games. To this end we introduce the angular policy gradient (APG), a stochastic policy gradient method for directional control. With the marginal policy gradients family of estimators we present a unified analysis of the variance reduction properties of APG and CAPG; our results provide a stronger guarantee than existing analyses for CAPG. Experimental results on a popular RTS game and a navigation task show that the APG estimator offers a substantial improvement over the standard policy gradient.

## code2seq: Generating Sequences from Structured Representations of Code

- Uri Alon, Shaked Brody, Omer Levy, Eran Yahav
- abstract@open-review(Poster): The ability to generate natural language sequences from source code snippets has a variety of applications such as code summarization, documentation, and retrieval. Sequence-to-sequence (seq2seq) models, adopted from neural machine translation (NMT), have achieved state-of-the-art performance on these tasks by treating source code as a sequence of tokens. We present code2seq: an alternative approach that leverages the syntactic structure of programming languages to better encode source code. Our model represents a code snippet as the set of compositional paths in its abstract syntax tree (AST) and uses attention to select the relevant paths while decoding. We demonstrate the effectiveness of our approach for two tasks, two programming languages, and four datasets of up to 16M examples. Our model significantly outperforms previous models that were specifically designed for programming languages, as well as general state-of-the-art NMT models. An interactive online demo of our model is available at http://code2seq.org. Our code, data and trained models are available at http://github.com/tech-srl/code2seq.

## ADef: an Iterative Algorithm to Construct Adversarial Deformations

- Rima Alaifari, Giovanni S. Alberti, Tandri Gauksson
- abstract@open-review(Poster): While deep neural networks have proven to be a powerful tool for many recognition and classification tasks, their stability properties are still not well understood. In the past, image classifiers have been shown to be vulnerable to so-called adversarial attacks, which are created by additively perturbing the correctly classified image. In this paper, we propose the ADef algorithm to construct a different kind of adversarial attack created by iteratively applying small deformations to the image, found through a gradient descent step. We demonstrate our results on MNIST with convolutional neural networks and on ImageNet with Inception-v3 and ResNet-101.

## Small nonlinearities in activation functions create bad local minima in neural networks

- Chulhee Yun, Suvrit Sra, Ali Jadbabaie
- abstract@open-review(Poster): We investigate the loss surface of neural networks. We prove that even for one-hidden-layer networks with "slightest" nonlinearity, the empirical risks have spurious local minima in most cases. Our results thus indicate that in general "no spurious local minim" is a property limited to deep linear networks, and insights obtained from linear networks may not be robust. Specifically, for ReLU(-like) networks we constructively prove that for almost all practical datasets there exist infinitely many local minima. We also present a counterexample for more general activations (sigmoid, tanh, arctan, ReLU, etc.), for which there exists a bad local minimum. Our results make the least restrictive assumptions relative to existing results on spurious local optima in neural networks. We complete our discussion by presenting a comprehensive characterization of global optimality for deep linear networks, which unifies other results on this topic.

## Visceral Machines: Risk-Aversion in Reinforcement Learning with Intrinsic Physiological Rewards

- Daniel McDuff, Ashish Kapoor
- abstract@open-review(Poster): As people learn to navigate the world, autonomic nervous system (e.g., ``fight or flight) responses provide intrinsic feedback about the potential consequence of action choices (e.g., becoming nervous when close to a cliff edge or driving fast around a bend.) Physiological changes are correlated with these biological preparations to protect one-self from danger. We present a novel approach to reinforcement learning that leverages a task-independent intrinsic reward function trained on peripheral pulse measurements that are correlated with human autonomic nervous system responses. Our hypothesis is that such reward functions can circumvent the challenges associated with sparse and skewed rewards in reinforcement learning settings and can help improve sample efficiency. We test this in a simulated driving environment and show that it can increase the speed of learning and reduce the number of collisions during the learning stage.

## ACCELERATING NONCONVEX LEARNING VIA REPLICA EXCHANGE LANGEVIN DIFFUSION

- Yi Chen, Jinglin Chen, Jing Dong, Jian Peng, Zhaoran Wang
- abstract@open-review(Poster): Langevin diffusion is a powerful method for nonconvex optimization, which enables the escape from local minima by injecting noise into the gradient. In particular, the temperature parameter controlling the noise level gives rise to a tradeoff between `global exploration'' and`local exploitation", which correspond to high and low temperatures. To attain the advantages of both regimes, we propose to use replica exchange, which swaps between two Langevin diffusions with different temperatures. We theoretically analyze the acceleration effect of replica exchange from two perspectives: (i) the convergence in $\chi^2$-divergence, and (ii) the large deviation principle. Such an acceleration effect allows us to

faster approach the global minima. Furthermore, by discretizing the replica exchange Langevin diffusion, we obtain a discrete-time algorithm. For such an algorithm, we quantify its discretization error in theory and demonstrate its acceleration effect in practice.

## The Singular Values of Convolutional Layers

- Hanie Sedghi, Vineet Gupta, Philip M. Long
- abstract@open-review(Poster): We characterize the singular values of the linear transformation associated with a standard 2D multi-channel convolutional layer, enabling their efficient computation. This characterization also leads to an algorithm for projecting a convolutional layer onto an operator-norm ball. We show that this is an effective regularizer; for example, it improves the test error of a deep residual network using batch normalization on CIFAR-10 from 6.2% to 5.3%.

## Improving Generalization and Stability of Generative Adversarial Networks

- Hoang Thanh-Tung, Truyen Tran, Svetha Venkatesh
- abstract@open-review(Poster): Generative Adversarial Networks (GANs) are one of the most popular tools for learning complex high dimensional distributions. However, generalization properties of GANs have not been well understood. In this paper, we analyze the generalization of GANs in practical settings. We show that discriminators trained on discrete datasets with the original GAN loss have poor generalization capability and do not approximate the theoretically optimal discriminator. We propose a zero-centered gradient penalty for improving the generalization of the discriminator by pushing it toward the optimal discriminator. The penalty guarantees the generalization and convergence of GANs. Experiments on synthetic and large scale datasets verify our theoretical analysis.

## GamePad: A Learning Environment for Theorem Proving

- Daniel Huang, Prafulla Dhariwal, Dawn Song, Ilya Sutskever
- abstract@open-review(Poster): In this paper, we introduce a system called GamePad that can be used to explore the application of machine learning methods to theorem proving in the Coq proof assistant. Interactive theorem provers such as Coq enable users to construct machine-checkable proofs in a step-by-step manner. Hence, they provide an opportunity to explore theorem proving with human supervision. We use GamePad to synthesize proofs for a simple algebraic rewrite problem and train baseline models for a formalization of the Feit-Thompson theorem. We address position evaluation (i.e., predict the number of proof steps left) and tactic prediction (i.e., predict the next proof step) tasks, which arise naturally in tactic-based theorem proving.

## Towards Understanding Regularization in Batch Normalization

- Ping Luo, Xinjiang Wang, Wenqi Shao, Zhanglin Peng
- abstract@open-review(Poster): Batch Normalization (BN) improves both convergence and generalization in training neural networks. This work understands these phenomena theoretically. We analyze BN by using a basic block of neural networks, consisting of a kernel layer, a BN layer, and a nonlinear activation function. This basic network helps us understand the impacts of BN in three aspects. First, by viewing BN as an implicit regularizer, BN can be decomposed into population normalization (PN) and gamma decay as an explicit regularization. Second, learning dynamics of BN and the regularization show that training converged with large maximum and effective learning rate. Third, generalization of BN is explored by using statistical mechanics. Experiments demonstrate that BN in convolutional neural networks share the same traits of regularization as the above analyses.

## Learning to Make Analogies by Contrasting Abstract Relational Structure

- Felix Hill, Adam Santoro, David Barrett, Ari Morcos, Timothy Lillicrap
- abstract@open-review(Poster): Analogical reasoning has been a principal focus of various waves of AI research. Analogy is particularly challenging for machines because it requires relational structures to be represented such that they can be flexibly applied across diverse domains of experience. Here, we study how analogical reasoning can be induced in neural networks that learn to perceive and reason about raw visual data. We find that the critical factor for inducing such a capacity is not an elaborate architecture, but rather, careful attention to the choice of data and the manner in which it is presented to the model. The most robust capacity for analogical reasoning is induced when networks learn analogies by contrasting abstract relational structures in their input domains, a training method that uses only the input data to force models to learn about important abstract features. Using this technique we demonstrate capacities for complex, visual and symbolic analogy making and generalisation in even the simplest neural network architectures.

## Attention, Learn to Solve Routing Problems!

- Wouter Kool, Herke van Hoof, Max Welling
- abstract@open-review(Poster): The recently presented idea to learn heuristics for combinatorial optimization problems is promising as it can save costly development. However, to push this idea towards practical implementation, we need better models and better ways of training. We contribute in both directions: we propose a model based on attention layers with benefits over the Pointer Network and we show how to train this model using REINFORCE with a simple baseline based on a deterministic greedy rollout, which we find is more efficient than using a value function. We significantly improve over recent learned heuristics for the Travelling Salesman Problem (TSP), getting close to optimal results for problems up to 100 nodes. With the same hyperparameters, we learn strong heuristics for two variants of the Vehicle Routing Problem (VRP), the Orienteering Problem (OP) and (a stochastic variant of) the Prize Collecting TSP (PCTSP), outperforming a wide range of baselines and getting results close to highly optimized and specialized algorithms.

## Critical Learning Periods in Deep Networks

- Alessandro Achille, Matteo Rovere, Stefano Soatto
- abstract@open-review(Poster): Similar to humans and animals, deep artificial neural networks exhibit critical periods during which a temporary stimulus deficit can impair the development of a skill. The extent of the impairment depends on the onset and length of the deficit window, as in animal models, and on the size of the neural network. Deficits that do not affect low-level statistics, such as vertical flipping of the images, have no lasting effect on performance and can be overcome with further training. To better understand this phenomenon, we use the Fisher Information of the weights to measure the effective connectivity between layers of a network during training. Counterintuitively, information rises rapidly in the early phases of training, and then decreases, preventing redistribution of information resources in a phenomenon we refer to as a loss of "Information Plasticity". Our analysis suggests that the first few epochs are critical for the creation of strong connections that are optimal relative to the input data distribution. Once such strong connections are created, they do not appear to change during additional training. These findings suggest that the initial learning transient, under-scrutinized compared to asymptotic behavior, plays a key role in determining the outcome of the training process. Our findings, combined with recent theoretical results in the literature, also suggest that forgetting (decrease of information in the weights) is critical to achieving invariance and disentanglement in representation learning. Finally, critical periods are not restricted to biological systems, but can emerge naturally in learning systems, whether biological or artificial, due to fundamental constrains arising from learning dynamics and information processing.

## Meta-Learning Probabilistic Inference for Prediction

- Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, Richard Turner
- abstract@open-review(Poster): This paper introduces a new framework for data efficient and versatile learning. Specifically: 1) We develop ML-PIP, a general framework for Meta-Learning approximate Probabilistic Inference for Prediction. ML-PIP extends existing probabilistic interpretations of meta-learning to cover a broad class of methods. 2) We introduce \Versa{}, an instance of the framework employing a flexible and versatile amortization network that takes few-shot learning datasets as inputs, with arbitrary numbers of shots, and outputs a distribution over task-specific parameters in a single forward pass. \Versa{} substitutes optimization at test time with forward passes through inference networks, amortizing the cost of inference and relieving the need for second derivatives during training. 3) We evaluate \Versa{} on benchmark datasets where the method sets new state-of-the-art results, and can handle arbitrary number of shots, and for classification, arbitrary numbers of classes at train and test time. The power of the approach is then demonstrated through a challenging few-shot ShapeNet view reconstruction task.

## Trellis Networks for Sequence Modeling

- Shaojie Bai, J. Zico Kolter, Vladlen Koltun
- abstract@open-review(Poster): We present trellis networks, a new architecture for sequence modeling. On the one hand, a trellis network is a temporal convolutional network with special structure, characterized by weight tying across depth and direct injection of the input into deep layers. On the other hand, we show that truncated recurrent networks are equivalent to trellis networks with special sparsity structure in their weight matrices. Thus trellis networks with general weight matrices generalize truncated recurrent networks. We leverage these connections to design high-performing trellis networks that absorb structural and algorithmic elements from both recurrent and convolutional models. Experiments demonstrate that trellis networks outperform the current state of the art methods on a variety of challenging benchmarks, including word-level language modeling and character-level language modeling tasks, and stress tests designed to evaluate long-term memory retention. The code is available at https://github.com/locuslab/trellisnet .

## Generative Code Modeling with Graphs

- Marc Brockschmidt, Miltiadis Allamanis, Alexander L. Gaunt, Oleksandr Polozov
- abstract@open-review(Poster): Generative models forsource code are an interesting structured prediction problem, requiring to reason about both hard syntactic and semantic constraints as well as about natural, likely programs. We present a novel model for this problem that uses a graph to represent the intermediate state of the generated output. Our model generates code by interleaving grammar-driven expansion steps with graph augmentation and neural message passing steps. An experimental evaluation shows that our new model can generate semantically meaningful expressions, outperforming a range of strong baselines.

## Learning Recurrent Binary/Ternary Weights

- Arash Ardakani, Zhengyun Ji, Sean C. Smithson, Brett H. Meyer, Warren J. Gross
- abstract@open-review(Poster): Recurrent neural networks (RNNs) have shown excellent performance in processing sequence data. However, they are both complex and memory intensive due to their recursive nature. These limitations make RNNs difficult to embed on mobile devices requiring real-time processes with limited hardware resources. To address the above issues, we introduce a method that can learn binary and ternary weights during the training phase to facilitate hardware implementations of RNNs. As a result, using this approach replaces all multiply-accumulate operations by simple accumulations, bringing significant benefits to custom hardware in terms of silicon area and power consumption. On the software side, we evaluate the performance (in terms of accuracy) of our method using long short-term memories (LSTMs) and gated recurrent units (GRUs) on various sequential models including sequence classification and language modeling. We demonstrate that our method achieves competitive results on the aforementioned tasks while using binary/ternary weights during the runtime. On the hardware side, we present custom hardware for accelerating the recurrent computations of LSTMs with binary/ternary weights. Ultimately, we show that LSTMs with binary/ternary weights can achieve up to 12x memory saving and 10x inference speedup compared to the full-precision hardware implementation design.

## Dynamically Unfolding Recurrent Restorer: A Moving Endpoint Control Method for Image Restoration

- Xiaoshuai Zhang, Yiping Lu, Jiaying Liu, Bin Dong
- abstract@open-review(Poster): In this paper, we propose a new control framework called the moving endpoint control to restore images corrupted by different degradation levels in one model. The proposed control problem contains a restoration dynamics which is modeled by an RNN. The moving endpoint, which is essentially the terminal time of the associated dynamics, is determined by a policy network. We call the proposed model the dynamically unfolding recurrent restorer (DURR). Numerical experiments show that DURR is able to achieve state-of-the-art performances on blind image denoising and JPEG image deblocking. Furthermore, DURR can well generalize to images with higher degradation levels that are not included in the training stage.

## Learning a SAT Solver from Single-Bit Supervision

- Daniel Selsam, Matthew Lamm, Benedikt B\"{u}nz, Percy Liang, Leonardo de Moura, David L. Dill
- abstract@open-review(Poster): We present NeuroSAT, a message passing neural network that learns to solve SAT problems after only being trained as a classifier to predict satisfiability. Although it is not competitive with state-of-the-art SAT solvers, NeuroSAT can solve problems that are substantially larger and more difficult than it ever saw during training by simply running for more iterations. Moreover, NeuroSAT generalizes to novel distributions; after training only on random SAT problems, at test time it can solve SAT problems encoding graph coloring, clique detection, dominating set, and vertex cover problems, all on a range of distributions over small random graphs.

## Optimal Transport Maps For Distribution Preserving Operations on Latent Spaces of Generative Models

- Eirikur Agustsson, Alexander Sage, Radu Timofte, Luc Van Gool
- abstract@open-review(Poster): Generative models such as Variational Auto Encoders (VAEs) and Generative Adversarial Networks (GANs) are typically trained for a fixed prior distribution in the latent space, such as uniform or Gaussian. After a trained model is obtained, one can sample the Generator in various forms for exploration and understanding, such as interpolating between two samples, sampling in the vicinity of a sample or exploring differences between a pair of samples applied to a third sample. However, the latent space operations commonly used in the literature so far induce a distribution mismatch between the resulting outputs and the prior distribution the model was trained on. Previous works have attempted to reduce this mismatch with heuristic modification to the operations or by changing the latent distribution and re-training models. In this paper, we propose a framework for modifying the latent space operations such that the distribution mismatch is fully eliminated. Our approach is based on optimal transport maps, which adapt the latent space operations such that they fully match the prior distribution, while minimally modifying the original operation. Our matched operations are readily obtained for the commonly used operations and distributions and require no adjustment to the training procedure.

## Efficient Lifelong Learning with A-GEM

- Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, Mohamed Elhoseiny
- abstract@open-review(Poster): In lifelong learning, the learner is presented with a sequence of tasks, incrementally building a data-driven prior which may be leveraged to speed up learning of a new task. In this work, we investigate the efficiency of current lifelong approaches, in terms of sample

complexity, computational and memory cost. Towards this end, we first introduce a new and a more realistic evaluation protocol, whereby learners observe each example only once and hyper-parameter selection is done on a small and disjoint set of tasks, which is not used for the actual learning experience and evaluation. Second, we introduce a new metric measuring how quickly a learner acquires a new skill. Third, we propose an improved version of GEM (Lopez-Paz & Ranzato, 2017), dubbed Averaged GEM (A-GEM), which enjoys the same or even better performance as GEM, while being almost as computationally and memory efficient as EWC (Kirkpatrick et al., 2016) and other regularization-based methods. Finally, we show that all algorithms including A-GEM can learn even more quickly if they are provided with task descriptors specifying the classification tasks under consideration. Our experiments on several standard lifelong learning benchmarks demonstrate that A-GEM has the best trade-off between accuracy and efficiency

## Meta-Learning For Stochastic Gradient MCMC

- Wenbo Gong, Yingzhen Li, José Miguel Hernández-Lobato
- abstract@open-review(Poster): Stochastic gradient Markov chain Monte Carlo (SG-MCMC) has become increasingly popular for simulating posterior samples in large-scale Bayesian modeling. However, existing SG-MCMC schemes are not tailored to any specific probabilistic model, even a simple modification of the underlying dynamical system requires significant physical intuition. This paper presents the first meta-learning algorithm that allows automated design for the underlying continuous dynamics of an SG-MCMC sampler. The learned sampler generalizes Hamiltonian dynamics with state-dependent drift and diffusion, enabling fast traversal and efficient exploration of energy landscapes. Experiments validate the proposed approach on Bayesian fully connected neural network, Bayesian convolutional neural network and Bayesian recurrent neural network tasks, showing that the learned sampler outperforms generic, hand-designed SG-MCMC algorithms, and generalizes to different datasets and larger architectures.

## Augmented Cyclic Adversarial Learning for Low Resource Domain Adaptation

- Ehsan Hosseini-Asl, Yingbo Zhou, Caiming Xiong, Richard Socher
- abstract@open-review(Poster): Training a model to perform a task typically requires a large amount of data from the domains in which the task will be applied. However, it is often the case that data are abundant in some domains but scarce in others. Domain adaptation deals with the challenge of adapting a model trained from a data-rich source domain to perform well in a data-poor target domain. In general, this requires learning plausible mappings between domains. CycleGAN is a powerful framework that efficiently learns to map inputs from one domain to another using adversarial training and a cycle-consistency constraint. However, the conventional approach of enforcing cycle-consistency via reconstruction may be overly restrictive in cases where one or more domains have limited training data. In this paper, we propose an augmented cyclic adversarial learning model that enforces the cycle-consistency constraint via an external task specific model, which encourages the preservation of task-relevant content as opposed to exact reconstruction. We explore digit classification in a low-resource setting in supervised, semi and unsupervised situation, as well as high resource unsupervised. In low-resource supervised setting, the results show that our approach improves absolute performance by 14% and 4% when adapting SVHN to MNIST and vice versa, respectively, which outperforms unsupervised domain adaptation methods that require high-resource unlabeled target domain. Moreover, using only few unsupervised target data, our approach can still outperforms many high-resource unsupervised models. Our model also outperforms on USPS to MNIST and synthetic digit to SVHN for high resource unsupervised adaptation. In speech domains, we similarly adopt a speech recognition model from each domain as the task specific model. Our approach improves absolute performance of speech recognition by 2% for female speakers in the TIMIT dataset, where the majority of training samples are from male voices.

## Off-Policy Evaluation and Learning from Logged Bandit Feedback: Error Reduction via Surrogate Policy

- Yuan Xie, Boyi Liu, Qiang Liu, Zhaoran Wang, Yuan Zhou, Jian Peng
- abstract@open-review(Poster): When learning from a batch of logged bandit feedback, the discrepancy between the policy to be learned and the off-policy training data imposes statistical and computational challenges. Unlike classical supervised learning and online learning settings, in batch contextual bandit learning, one only has access to a collection of logged feedback from the actions taken by a historical policy, and expect to learn a policy that takes good actions in possibly unseen contexts. Such a batch learning setting is ubiquitous in online and interactive systems, such as ad platforms and recommendation systems. Existing approaches based on inverse propensity weights, such as Inverse Propensity Scoring (IPS) and Policy Optimizer for Exponential Models (POEM), enjoy unbiasedness but often suffer from large mean squared error. In this work, we introduce a new approach named Maximum Likelihood Inverse Propensity Scoring (MLIPS) for batch learning from logged bandit feedback. Instead of using the given historical policy as the proposal in inverse propensity weights, we estimate a maximum likelihood surrogate policy based on the logged action-context pairs, and then use this surrogate policy as the proposal. We prove that MLIPS is asymptotically unbiased, and moreover, has a smaller nonasymptotic mean squared error than IPS. Such an error reduction phenomenon is somewhat surprising as the estimated surrogate policy is less accurate than the given historical policy. Results on multi-label classification problems and a large-scale ad placement dataset demonstrate the empirical effectiveness of MLIPS. Furthermore, the proposed surrogate policy technique is complementary to existing error reduction techniques, and when combined, is able to consistently boost the performance of several widely used approaches.

## Double Viterbi: Weight Encoding for High Compression Ratio and Fast On-Chip Reconstruction for Deep Neural Network

- Daehyun Ahn, Dongsoo Lee, Taesu Kim, Jae-Joon Kim
- abstract@open-review(Poster): Weight pruning has been introduced as an efficient model compression technique. Even though pruning removes significant amount of weights in a network, memory requirement reduction was limited since conventional sparse matrix formats require significant amount of memory to store index-related information. Moreover, computations associated with such sparse matrix formats are slow because sequential sparse matrix decoding process does not utilize highly parallel computing systems efficiently. As an attempt to compress index information while keeping the decoding process parallelizable, Viterbi-based pruning was suggested. Decoding non-zero weights, however, is still sequential in Viterbi-based pruning. In this paper, we propose a new sparse matrix format in order to enable a highly parallel decoding process of the entire sparse matrix. The proposed sparse matrix is constructed by combining pruning and weight quantization. For the latest RNN models on PTB and WikiText-2 corpus, LSTM parameter storage requirement is compressed 19x using the proposed sparse matrix format compared to the baseline model. Compressed weight and indices can be reconstructed into a dense matrix fast using Viterbi encoders. Simulation results show that the proposed scheme can feed parameters to processing elements 20 % to 106 % faster than the case where the dense matrix values directly come from DRAM.

## Graph Wavelet Neural Network

- Bingbing Xu, Huawei Shen, Qi Cao, Yunqi Qiu, Xueqi Cheng
- abstract@open-review(Poster): We present graph wavelet neural network (GWNN), a novel graph convolutional neural network (CNN), leveraging graph wavelet transform to address the shortcomings of previous spectral graph CNN methods that depend on graph Fourier transform. Different from graph Fourier transform, graph wavelet transform can be obtained via a fast algorithm without requiring matrix eigendecomposition with high computational cost. Moreover, graph wavelets are sparse and localized in vertex domain, offering high efficiency and good interpretability for graph convolution. The proposed GWNN significantly outperforms previous spectral graph CNNs in the task of graph-based semi-supervised classification on three benchmark datasets: Cora, Citeseer and Pubmed.

## The Unusual Effectiveness of Averaging in GAN Training

- Yasin Yaz{\i}c{\i}, Chuan-Sheng Foo, Stefan Winkler, Kim-Hui Yap, Georgios Piliouras, Vijay Chandrasekhar

- abstract@open-review(Poster): We examine two different techniques for parameter averaging in GAN training. Moving Average (MA) computes the time-average of parameters, whereas Exponential Moving Average (EMA) computes an exponentially discounted sum. Whilst MA is known to lead to convergence in bilinear settings, we provide the -- to our knowledge -- first theoretical arguments in support of EMA. We show that EMA converges to limit cycles around the equilibrium with vanishing amplitude as the discount parameter approaches one for simple bilinear games and also enhances the stability of general GAN training. We establish experimentally that both techniques are strikingly effective in the non-convex-concave GAN setting as well. Both improve inception and FID scores on different architectures and for different GAN objectives. We provide comprehensive experimental results across a range of datasets -- mixture of Gaussians, CIFAR-10, STL-10, CelebA and ImageNet -- to demonstrate its effectiveness. We achieve state-of-the-art results on CIFAR-10 and produce clean CelebA face images.\footnote{~The code is available at \url{https://github.com/yasinyazici/EMA_GAN}}

## Evaluating Robustness of Neural Networks with Mixed Integer Programming

- Vincent Tjeng, Kai Y. Xiao, Russ Tedrake
- abstract@open-review(Poster): Neural networks trained only to optimize for training accuracy can often be fooled by adversarial examples --- slightly perturbed inputs misclassified with high confidence. Verification of networks enables us to gauge their vulnerability to such adversarial examples. We formulate verification of piecewise-linear neural networks as a mixed integer program. On a representative task of finding minimum adversarial distortions, our verifier is two to three orders of magnitude quicker than the state-of-the-art. We achieve this computational speedup via tight formulations for non-linearities, as well as a novel presolve algorithm that makes full use of all information available. The computational speedup allows us to verify properties on convolutional and residual networks with over 100,000 ReLUs --- several orders of magnitude more than networks previously verified by any complete verifier. In particular, we determine for the first time the exact adversarial accuracy of an MNIST classifier to perturbations with bounded l-∞ norm ε=0.1: for this classifier, we find an adversarial example for 4.38% of samples, and a certificate of robustness to norm-bounded perturbations for the remainder. Across all robust training procedures and network architectures considered, and for both the MNIST and CIFAR-10 datasets, we are able to certify more samples than the state-of-the-art and find more adversarial examples than a strong first-order attack.

## Towards the first adversarially robust neural network model on MNIST

- Lukas Schott, Jonas Rauber, Matthias Bethge, Wieland Brendel
- abstract@open-review(Poster): Despite much effort, deep neural networks remain highly susceptible to tiny input perturbations and even for MNIST, one of the most common toy datasets in computer vision, no neural network model exists for which adversarial perturbations are large and make semantic sense to humans. We show that even the widely recognized and by far most successful L-inf defense by Madry et~al. (1) has lower L0 robustness than undefended networks and still highly susceptible to L2 perturbations, (2) classifies unrecognizable images with high certainty, (3) performs not much better than simple input binarization and (4) features adversarial perturbations that make little sense to humans. These results suggest that MNIST is far from being solved in terms of adversarial robustness. We present a novel robust classification model that performs analysis by synthesis using learned class-conditional data distributions. We derive bounds on the robustness and go to great length to empirically evaluate our model using maximally effective adversarial attacks by (a) applying decision-based, score-based, gradient-based and transfer-based attacks for several different Lp norms, (b) by designing a new attack that exploits the structure of our defended model and (c) by devising a novel decision-based attack that seeks to minimize the number of perturbed pixels (L0). The results suggest that our approach yields state-of-the-art robustness on MNIST against L0, L2 and L-inf perturbations and we demonstrate that most adversarial examples are strongly perturbed towards the perceptual boundary between the original and the adversarial class.

## On the Turing Completeness of Modern Neural Network Architectures

- Jorge Pérez, Javier Marinković, Pablo Barceló
- abstract@open-review(Poster): Alternatives to recurrent neural networks, in particular, architectures based on attention or convolutions, have been gaining momentum for processing input sequences. In spite of their relevance, the computational properties of these alternatives have not yet been fully explored. We study the computational power of two of the most paradigmatic architectures exemplifying these mechanisms: the Transformer (Vaswani et al., 2017) and the Neural GPU (Kaiser & Sutskever, 2016). We show both models to be Turing complete exclusively based on their capacity to compute and access internal dense representations of the data. In particular, neither the Transformer nor the Neural GPU requires access to an external memory to become Turing complete. Our study also reveals some minimal sets of elements needed to obtain these completeness results.

## Adaptive Posterior Learning: few-shot learning with a surprise-based memory module

- Tiago Ramalho, Marta Garnelo
- abstract@open-review(Poster): The ability to generalize quickly from few observations is crucial for intelligent systems. In this paper we introduce APL, an algorithm that approximates probability distributions by remembering the most surprising observations it has encountered. These past observations are recalled from an external memory module and processed by a decoder network that can combine information from different memory slots to generalize beyond direct recall. We show this algorithm can perform as well as state of the art baselines on few-shot classification benchmarks with a smaller memory footprint. In addition, its memory compression allows it to scale to thousands of unknown labels. Finally, we introduce a meta-learning reasoning task which is more challenging than direct classification. In this setting, APL is able to generalize with fewer than one example per class via deductive reasoning.

## A Closer Look at Deep Learning Heuristics: Learning rate restarts, Warmup and Distillation

- Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, Richard Socher
- abstract@open-review(Poster): The convergence rate and final performance of common deep learning models have significantly benefited from recently proposed heuristics such as learning rate schedules, knowledge distillation, skip connections and normalization layers. In the absence of theoretical underpinnings, controlled experiments aimed at explaining the efficacy of these strategies can aid our understanding of deep learning landscapes and the training dynamics. Existing approaches for empirical analysis rely on tools of linear interpolation and visualizations with dimensionality reduction, each with their limitations. Instead, we revisit the empirical analysis of heuristics through the lens of recently proposed methods for loss surface and representation analysis, viz. mode connectivity and canonical correlation analysis (CCA), and hypothesize reasons why the heuristics succeed. In particular, we explore knowledge distillation and learning rate heuristics of (cosine) restarts and warmup using mode connectivity and CCA. Our empirical analysis suggests that: (a) the reasons often quoted for the success of cosine annealing are not evidenced in practice; (b) that the effect of learning rate warmup is to prevent the deeper layers from creating training instability; and (c) that the latent knowledge shared by the teacher is primarily disbursed in the deeper layers.

## Coarse-grain Fine-grain Coattention Network for Multi-evidence Question Answering

- Victor Zhong, Caiming Xiong, Nitish Shirish Keskar, Richard Socher
- abstract@open-review(Poster): End-to-end neural models have made significant progress in question answering, however recent studies show that these models implicitly assume that the answer and evidence appear close together in a single document. In this work, we propose the Coarse-grain Fine-grain Coattention Network (CFC), a new question answering model that combines information from evidence across multiple documents. The CFC consists of a coarse-grain module that interprets documents with respect to the query then finds a relevant answer, and a fine-grain module which scores each candidate answer by comparing its occurrences across all of the documents with the query. We design these modules using hierarchies of coattention and self-

attention, which learn to emphasize different parts of the input. On the Qangaroo WikiHop multi-evidence question answering task, the CFC obtains a new state-of-the-art result of 70.6% on the blind test set, outperforming the previous best by 3% accuracy despite not using pretrained contextual encoders.

## Near-Optimal Representation Learning for Hierarchical Reinforcement Learning

- Ofir Nachum, Shixiang Gu, Honglak Lee, Sergey Levine
- abstract@open-review(Poster): We study the problem of representation learning in goal-conditioned hierarchical reinforcement learning. In such hierarchical structures, a higher-level controller solves tasks by iteratively communicating goals which a lower-level policy is trained to reach. Accordingly, the choice of representation -- the mapping of observation space to goal space -- is crucial. To study this problem, we develop a notion of sub-optimality of a representation, defined in terms of expected reward of the optimal hierarchical policy using this representation. We derive expressions which bound the sub-optimality and show how these expressions can be translated to representation learning objectives which may be optimized in practice. Results on a number of difficult continuous-control tasks show that our approach to representation learning yields qualitatively better representations as well as quantitatively better hierarchical policies, compared to existing methods.

## Execution-Guided Neural Program Synthesis

- Xinyun Chen, Chang Liu, Dawn Song
- abstract@open-review(Poster): Neural program synthesis from input-output examples has attracted an increasing interest from both the machine learning and the programming language community. Most existing neural program synthesis approaches employ an encoder-decoder architecture, which uses an encoder to compute the embedding of the given input-output examples, as well as a decoder to generate the program from the embedding following a given syntax. Although such approaches achieve a reasonable performance on simple tasks such as FlashFill, on more complex tasks such as Karel, the state-of-the-art approach can only achieve an accuracy of around 77%. We observe that the main drawback of existing approaches is that the semantic information is greatly under-utilized. In this work, we propose two simple yet principled techniques to better leverage the semantic information, which are execution-guided synthesis and synthesizer ensemble. These techniques are general enough to be combined with any existing encoder-decoder-style neural program synthesizer. Applying our techniques to the Karel dataset, we can boost the accuracy from around 77% to more than 90%.

## Imposing Category Trees Onto Word-Embeddings Using A Geometric Construction

- Tiansi Dong, Chrisitan Bauckhage, Hailong Jin, Juanzi Li, Olaf Cremers, Daniel Speicher, Armin B. Cremers, Joerg Zimmermann
- abstract@open-review(Poster): We present a novel method to precisely impose tree-structured category information onto word-embeddings, resulting in ball embeddings in higher dimensional spaces (N-balls for short). Inclusion relations among N-balls implicitly encode subordinate relations among categories. The similarity measurement in terms of the cosine function is enriched by category information. Using a geometric construction method instead of back-propagation, we create large N-ball embeddings that satisfy two conditions: (1) category trees are precisely imposed onto word embeddings at zero energy cost; (2) pre-trained word embeddings are well preserved. A new benchmark data set is created for validating the category of unknown words. Experiments show that N-ball embeddings, carrying category information, significantly outperform word embeddings in the test of nearest neighborhoods, and demonstrate surprisingly good performance in validating categories of unknown words. Source codes and data-sets are free for public access \url{https://github.com/gnodisnait/nball4tree.git} and \url{https://github.com/gnodisnait/bp94nball.git}.

## ROBUST ESTIMATION VIA GENERATIVE ADVERSARIAL NETWORKS

- Chao GAO, jiyi LIU, Yuan YAO, Weizhi ZHU
- abstract@open-review(Poster): Robust estimation under Huber's $\epsilon$-contamination model has become an important topic in statistics and theoretical computer science. Rate-optimal procedures such as Tukey's median and other estimators based on statistical depth functions are impractical because of their computational intractability. In this paper, we establish an intriguing connection between f-GANs and various depth functions through the lens of f-Learning. Similar to the derivation of f-GAN, we show that these depth functions that lead to rate-optimal robust estimators can all be viewed as variational lower bounds of the total variation distance in the framework of f-Learning. This connection opens the door of computing robust estimators using tools developed for training GANs. In particular, we show that a JS-GAN that uses a neural network discriminator with at least one hidden layer is able to achieve the minimax rate of robust mean estimation under Huber's $\epsilon$-contamination model. Interestingly, the hidden layers of the neural net structure in the discriminator class are shown to be necessary for robust estimation.

## Woulda, Coulda, Shoulda: Counterfactually-Guided Policy Search

- Lars Buesing, Theophane Weber, Yori Zwols, Nicolas Heess, Sebastien Racaniere, Arthur Guez, Jean-Baptiste Lespiau
- abstract@open-review(Poster): Learning policies on data synthesized by models can in principle quench the thirst of reinforcement learning algorithms for large amounts of real experience, which is often costly to acquire. However, simulating plausible experience de novo is a hard problem for many complex environments, often resulting in biases for model-based policy evaluation and search. Instead of de novo synthesis of data, here we assume logged, real experience and model alternative outcomes of this experience under counterfactual actions, i.e. actions that were not actually taken. Based on this, we propose the Counterfactually-Guided Policy Search (CF-GPS) algorithm for learning policies in POMDPs from off-policy experience. It leverages structural causal models for counterfactual evaluation of arbitrary policies on individual off-policy episodes. CF-GPS can improve on vanilla model-based RL algorithms by making use of available logged data to de-bias model predictions. In contrast to off-policy algorithms based on Importance Sampling which re-weight data, CF-GPS leverages a model to explicitly consider alternative outcomes, allowing the algorithm to make better use of experience data. We find empirically that these advantages translate into improved policy evaluation and search results on a non-trivial grid-world task. Finally, we show that CF-GPS generalizes the previously proposed Guided Policy Search and that reparameterization-based algorithms such Stochastic Value Gradient can be interpreted as counterfactual methods.

## Robust Conditional Generative Adversarial Networks

- Grigorios G. Chrysos, Jean Kossaifi, Stefanos Zafeiriou
- abstract@open-review(Poster): Conditional generative adversarial networks (cGAN) have led to large improvements in the task of conditional image generation, which lies at the heart of computer vision. The major focus so far has been on performance improvement, while there has been little effort in making cGAN more robust to noise. The regression (of the generator) might lead to arbitrarily large errors in the output, which makes cGAN unreliable for real-world applications. In this work, we introduce a novel conditional GAN model, called RoCGAN, which leverages structure in the target space of the model to address the issue. Our model augments the generator with an unsupervised pathway, which promotes the outputs of the generator to span the target manifold even in the presence of intense noise. We prove that RoCGAN share similar theoretical properties as GAN and experimentally verify that our model outperforms existing state-of-the-art cGAN architectures by a large margin in a variety of domains including images from natural scenes and faces.

## Attentive Neural Processes

- Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, Yee Whye Teh

- abstract@open-review(Poster): Neural Processes (NPs) (Garnelo et al., 2018) approach regression by learning to map a context set of observed input-output pairs to a distribution over regression functions. Each function models the distribution of the output given an input, conditioned on the context. NPs have the benefit of fitting observed data efficiently with linear complexity in the number of context input-output pairs, and can learn a wide family of conditional distributions; they learn predictive distributions conditioned on context sets of arbitrary size. Nonetheless, we show that NPs suffer a fundamental drawback of underfitting, giving inaccurate predictions at the inputs of the observed data they condition on. We address this issue by incorporating attention into NPs, allowing each input location to attend to the relevant context points for the prediction. We show that this greatly improves the accuracy of predictions, results in noticeably faster training, and expands the range of functions that can be modelled.

## Visual Reasoning by Progressive Module Networks

- Seung Wook Kim, Makarand Tapaswi, Sanja Fidler
- abstract@open-review(Poster): Humans learn to solve tasks of increasing complexity by building on top of previously acquired knowledge. Typically, there exists a natural progression in the tasks that we learn – most do not require completely independent solutions, but can be broken down into simpler subtasks. We propose to represent a solver for each task as a neural module that calls existing modules (solvers for simpler tasks) in a functional program-like manner. Lower modules are a black box to the calling module, and communicate only via a query and an output. Thus, a module for a new task learns to query existing modules and composes their outputs in order to produce its own output. Our model effectively combines previous skill-sets, does not suffer from forgetting, and is fully differentiable. We test our model in learning a set of visual reasoning tasks, and demonstrate improved performances in all tasks by learning progressively. By evaluating the reasoning process using human judges, we show that our model is more interpretable than an attention-based baseline.

## Hindsight policy gradients

- Paulo Rauber, Avinash Ummadisingu, Filipe Mutz, Jürgen Schmidhuber
- abstract@open-review(Poster): A reinforcement learning agent that needs to pursue different goals across episodes requires a goal-conditional policy. In addition to their potential to generalize desirable behavior to unseen goals, such policies may also enable higher-level planning based on subgoals. In sparse-reward environments, the capacity to exploit information about the degree to which an arbitrary goal has been achieved while another goal was intended appears crucial to enable sample efficient learning. However, reinforcement learning agents have only recently been endowed with such capacity for hindsight. In this paper, we demonstrate how hindsight can be introduced to policy gradient methods, generalizing this idea to a broad class of successful algorithms. Our experiments on a diverse selection of sparse-reward environments show that hindsight leads to a remarkable increase in sample efficiency.

## LeMoNADe: Learned Motif and Neuronal Assembly Detection in calcium imaging videos

- Elke Kirschbaum, Manuel Haußmann, Steffen Wolf, Hannah Sonntag, Justus Schneider, Shehabeldin Elzoheiry, Oliver Kann, Daniel Durstewitz, Fred A Hamprecht
- abstract@open-review(Poster): Neuronal assemblies, loosely defined as subsets of neurons with reoccurring spatio-temporally coordinated activation patterns, or "motifs", are thought to be building blocks of neural representations and information processing. We here propose LeMoNADe, a new exploratory data analysis method that facilitates hunting for motifs in calcium imaging videos, the dominant microscopic functional imaging modality in neurophysiology. Our nonparametric method extracts motifs directly from videos, bypassing the difficult intermediate step of spike extraction. Our technique augments variational autoencoders with a discrete stochastic node, and we show in detail how a differentiable reparametrization and relaxation can be used. An evaluation on simulated data, with available ground truth, reveals excellent quantitative performance. In real video data acquired from brain slices, with no ground truth available, LeMoNADe uncovers nontrivial candidate motifs that can help generate hypotheses for more focused biological investigations.

## Visual Explanation by Interpretation: Improving Visual Feedback Capabilities of Deep Neural Networks

- Jose Oramas, Kaili Wang, Tinne Tuytelaars
- abstract@open-review(Poster): Visual Interpretation and explanation of deep models is critical towards wide adoption of systems that rely on them. In this paper, we propose a novel scheme for both interpretation as well as explanation in which, given a pretrained model, we automatically identify internal features relevant for the set of classes considered by the model, without relying on additional annotations. We interpret the model through average visualizations of this reduced set of features. Then, at test time, we explain the network prediction by accompanying the predicted class label with supporting visualizations derived from the identified features. In addition, we propose a method to address the artifacts introduced by strided operations in deconvNet-based visualizations. Moreover, we introduce an8Flower , a dataset specifically designed for objective quantitative evaluation of methods for visual explanation. Experiments on the MNIST , ILSVRC 12, Fashion 144k and an8Flower datasets show that our method produces detailed explanations with good coverage of relevant features of the classes of interest.

## Max-MIG: an Information Theoretic Approach for Joint Learning from Crowds

- Peng Cao, *Yilun Xu*, Yuqing Kong, Yizhou Wang
- abstract@open-review(Poster): Eliciting labels from crowds is a potential way to obtain large labeled data. Despite a variety of methods developed for learning from crowds, a key challenge remains unsolved: \emph{learning from crowds without knowing the information structure among the crowds a priori, when some people of the crowds make highly correlated mistakes and some of them label effortlessly (e.g. randomly)}. We propose an information theoretic approach, Max-MIG, for joint learning from crowds, with a common assumption: the crowdsourced labels and the data are independent conditioning on the ground truth. Max-MIG simultaneously aggregates the crowdsourced labels and learns an accurate data classifier. Furthermore, we devise an accurate data-crowds forecaster that employs both the data and the crowdsourced labels to forecast the ground truth. To the best of our knowledge, this is the first algorithm that solves the aforementioned challenge of learning from crowds. In addition to the theoretical validation, we also empirically show that our algorithm achieves the new state-of-the-art results in most settings, including the real-world data, and is the first algorithm that is robust to various information structures. Codes are available at https://github.com/Newbeeer/Max-MIG .

## Hierarchical Visuomotor Control of Humanoids

- Josh Merel, Arun Ahuja, Vu Pham, Saran Tunyasuvunakool, Siqi Liu, Dhruva Tirumala, Nicolas Heess, Greg Wayne
- abstract@open-review(Poster): We aim to build complex humanoid agents that integrate perception, motor control, and memory. In this work, we partly factor this problem into low-level motor control from proprioception and high-level coordination of the low-level skills informed by vision. We develop an architecture capable of surprisingly flexible, task-directed motor control of a relatively high-DoF humanoid body by combining pre-training of low-level motor controllers with a high-level, task-focused controller that switches among low-level sub-policies. The resulting system is able to control a physically-simulated humanoid body to solve tasks that require coupling visual perception from an unstabilized egocentric RGB camera during locomotion in the environment. Supplementary video link: https://youtu.be/fBoir7PNxPk

## Function Space Particle Optimization for Bayesian Neural Networks

- Ziyu Wang, Tongzheng Ren, Jun Zhu, Bo Zhang
- abstract@open-review(Poster): While Bayesian neural networks (BNNs) have drawn increasing attention, their posterior inference remains challenging, due to the high-dimensional and over-parameterized nature. To address this issue, several highly flexible and scalable variational inference procedures based on the idea of particle optimization have been proposed. These methods directly optimize a set of particles to approximate the target posterior. However, their application to BNNs often yields sub-optimal performance, as such methods have a particular failure mode on over-parameterized models. In this paper, we propose to solve this issue by performing particle optimization directly in the space of regression functions. We demonstrate through extensive experiments that our method successfully overcomes this issue, and outperforms strong baselines in a variety of tasks including prediction, defense against adversarial examples, and reinforcement learning.

## Feed-forward Propagation in Probabilistic Neural Networks with Categorical and Max Layers

- Alexander Shekhovtsov, Boris Flach
- abstract@open-review(Poster): Probabilistic Neural Networks deal with various sources of stochasticity: input noise, dropout, stochastic neurons, parameter uncertainties modeled as random variables, etc. In this paper we revisit a feed-forward propagation approach that allows one to estimate for each neuron its mean and variance w.r.t. all mentioned sources of stochasticity. In contrast, standard NNs propagate only point estimates, discarding the uncertainty. Methods propagating also the variance have been proposed by several authors in different context. The view presented here attempts to clarify the assumptions and derivation behind such methods, relate them to classical NNs and broaden their scope of applicability. The main technical contributions are new approximations for the distributions of argmax and max-related transforms, which allow for fully analytic uncertainty propagation in networks with softmax and max-pooling layers as well as leaky ReLU activations. We evaluate the accuracy of the approximation and suggest a simple calibration. Applying the method to networks with dropout allows for faster training and gives improved test likelihoods without the need of sampling.

## Hierarchical Reinforcement Learning via Advantage-Weighted Information Maximization

- Takayuki Osa, Voot Tangkaratt, Masashi Sugiyama
- abstract@open-review(Poster): Real-world tasks are often highly structured. Hierarchical reinforcement learning (HRL) has attracted research interest as an approach for leveraging the hierarchical structure of a given task in reinforcement learning (RL). However, identifying the hierarchical policy structure that enhances the performance of RL is not a trivial task. In this paper, we propose an HRL method that learns a latent variable of a hierarchical policy using mutual information maximization. Our approach can be interpreted as a way to learn a discrete and latent representation of the state-action space. To learn option policies that correspond to modes of the advantage function, we introduce advantage-weighted importance sampling.
  In our HRL method, the gating policy learns to select option policies based on an option-value function, and these option policies are optimized based on the deterministic policy gradient method. This framework is derived by leveraging the analogy between a monolithic policy in standard RL and a hierarchical policy in HRL by using a deterministic option policy. Experimental results indicate that our HRL approach can learn a diversity of options and that it can enhance the performance of RL in continuous control tasks.

## Large-Scale Study of Curiosity-Driven Learning

- Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, Alexei A. Efros
- abstract@open-review(Poster): Reinforcement learning algorithms rely on carefully engineered rewards from the environment that are extrinsic to the agent. However, annotating each environment with hand-designed, dense rewards is difficult and not scalable, motivating the need for developing reward functions that are intrinsic to the agent. Curiosity is such intrinsic reward function which uses prediction error as a reward signal. In this paper: (a) We perform the first large-scale study of purely curiosity-driven learning, i.e. {\em without any extrinsic rewards}, across $54$ standard benchmark environments, including the Atari game suite. Our results show surprisingly good performance as well as a high degree of alignment between the intrinsic curiosity objective and the hand-designed extrinsic rewards of many games. (b) We investigate the effect of using different feature spaces for computing prediction error and show that random features are sufficient for many popular RL game benchmarks, but learned features appear to generalize better (e.g. to novel game levels in Super Mario Bros.). (c) We demonstrate limitations of the prediction-based rewards in stochastic setups. Game-play videos and code are at https://doubleblindsupplementary.github.io/large-curiosity/.

## StrokeNet: A Neural Painting Environment

- Ningyuan Zheng, Yifan Jiang, Dingjiang Huang
- abstract@open-review(Poster): We've seen tremendous success of image generating models these years. Generating images through a neural network is usually pixel-based, which is fundamentally different from how humans create artwork using brushes. To imitate human drawing, interactions between the environment and the agent is required to allow trials. However, the environment is usually non-differentiable, leading to slow convergence and massive computation. In this paper we try to address the discrete nature of software environment with an intermediate, differentiable simulation. We present StrokeNet, a novel model where the agent is trained upon a well-crafted neural approximation of the painting environment. With this approach, our agent was able to learn to write characters such as MNIST digits faster than reinforcement learning approaches in an unsupervised manner. Our primary contribution is the neural simulation of a real-world environment. Furthermore, the agent trained with the emulated environment is able to directly transfer its skills to real-world software.

## DialogWAE: Multimodal Response Generation with Conditional Wasserstein Auto-Encoder

- Xiaodong Gu, Kyunghyun Cho, Jung-Woo Ha, Sunghun Kim
- abstract@open-review(Poster): Variational autoencoders (VAEs) have shown a promise in data-driven conversation modeling. However, most VAE conversation models match the approximate posterior distribution over the latent variables to a simple prior such as standard normal distribution, thereby restricting the generated responses to a relatively simple (e.g., single-modal) scope. In this paper, we propose DialogWAE, a conditional Wasserstein autoencoder (WAE) specially designed for dialogue modeling. Unlike VAEs that impose a simple distribution over the latent variables, DialogWAE models the distribution of data by training a GAN within the latent variable space. Specifically, our model samples from the prior and posterior distributions over the latent variables by transforming context-dependent random noise using neural networks and minimizes the Wasserstein distance between the two distributions. We further develop a Gaussian mixture prior network to enrich the latent space. Experiments on two popular datasets show that DialogWAE outperforms the state-of-the-art approaches in generating more coherent, informative and diverse responses.

## Quaternion Recurrent Neural Networks

- Titouan Parcollet, Mirco Ravanelli, Mohamed Morchid, Georges Linarès, Chiheb Trabelsi, Renato De Mori, Yoshua Bengio
- abstract@open-review(Poster): Recurrent neural networks (RNNs) are powerful architectures to model sequential data, due to their capability to learn short and long-term dependencies between the basic elements of a sequence. Nonetheless, popular tasks such as speech or images recognition, involve multi-dimensional input features that are characterized by strong internal dependencies between the dimensions of the input vector. We propose a novel quaternion recurrent neural network (QRNN), alongside with a quaternion long-short term memory neural network (QLSTM), that take into account both the external relations and these internal structural dependencies with the quaternion algebra. Similarly to capsules, quaternions allow the QRNN to code internal dependencies by composing and processing multidimensional features as single entities, while the recurrent operation reveals correlations between the elements composing the sequence. We show that both QRNN and QLSTM achieve better performances than RNN and LSTM in a realistic application of automatic speech recognition. Finally, we show that QRNN and QLSTM reduce by a maximum factor of 3.3x the number of free

parameters needed, compared to real-valued RNNs and LSTMs to reach better results, leading to a more compact representation of the relevant information.

## Reward Constrained Policy Optimization

- Chen Tessler, Daniel J. Mankowitz, Shie Mannor
- abstract@open-review(Poster): Solving tasks in Reinforcement Learning is no easy feat. As the goal of the agent is to maximize the accumulated reward, it often learns to exploit loopholes and misspecifications in the reward signal resulting in unwanted behavior. While constraints may solve this issue, there is no closed form solution for general constraints. In this work we present a novel multi-timescale approach for constrained policy optimization, called `Reward Constrained Policy Optimization' (RCPO), which uses an alternative penalty signal to guide the policy towards a constraint satisfying one. We prove the convergence of our approach and provide empirical evidence of its ability to train constraint satisfying policies.

## Learning Latent Superstructures in Variational Autoencoders for Deep Multidimensional Clustering

- Xiaopeng Li, Zhourong Chen, Leonard K. M. Poon, Nevin L. Zhang
- abstract@open-review(Poster): We investigate a variant of variational autoencoders where there is a superstructure of discrete latent variables on top of the latent features. In general, our superstructure is a tree structure of multiple super latent variables and it is automatically learned from data. When there is only one latent variable in the superstructure, our model reduces to one that assumes the latent features to be generated from a Gaussian mixture model. We call our model the latent tree variational autoencoder (LTVAE). Whereas previous deep learning methods for clustering produce only one partition of data, LTVAE produces multiple partitions of data, each being given by one super latent variable. This is desirable because high dimensional data usually have many different natural facets and can be meaningfully partitioned in multiple ways.

## Variational Smoothing in Recurrent Neural Network Language Models

- Lingpeng Kong, Gabor Melis, Wang Ling, Lei Yu, Dani Yogatama
- abstract@open-review(Poster): We present a new theoretical perspective of data noising in recurrent neural network language models (Xie et al., 2017). We show that each variant of data noising is an instance of Bayesian recurrent neural networks with a particular variational distribution (i.e., a mixture of Gaussians whose weights depend on statistics derived from the corpus such as the unigram distribution). We use this insight to propose a more principled method to apply at prediction time and propose natural extensions to data noising under the variational framework. In particular, we propose variational smoothing with tied input and output embedding matrices and an element-wise variational smoothing method. We empirically verify our analysis on two benchmark language modeling datasets and demonstrate performance improvements over existing data noising methods.

## Initialized Equilibrium Propagation for Backprop-Free Training

- Peter O'Connor, Efstratios Gavves, Max Welling
- abstract@open-review(Poster): Deep neural networks are almost universally trained with reverse-mode automatic differentiation (a.k.a. backpropagation). Biological networks, on the other hand, appear to lack any mechanism for sending gradients back to their input neurons, and thus cannot be learning in this way. In response to this, Scellier & Bengio (2017) proposed Equilibrium Propagation - a method for gradient-based train- ing of neural networks which uses only local learning rules and, crucially, does not rely on neurons having a mechanism for back-propagating an error gradient. Equilibrium propagation, however, has a major practical limitation: inference involves doing an iterative optimization of neural activations to find a fixed-point, and the number of steps required to closely approximate this fixed point scales poorly with the depth of the network. In response to this problem, we propose Initialized Equilibrium Propagation, which trains a feedforward network to initialize the iterative inference procedure for Equilibrium propagation. This feed-forward network learns to approximate the state of the fixed-point using a local learning rule. After training, we can simply use this initializing network for inference, resulting in a learned feedforward network. Our experiments show that this network appears to work as well or better than the original version of Equilibrium propagation. This shows how we might go about training deep networks without using backpropagation.

## Identifying and Controlling Important Neurons in Neural Machine Translation

- Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, James Glass
- abstract@open-review(Poster): Neural machine translation (NMT) models learn representations containing substantial linguistic information. However, it is not clear if such information is fully distributed or if some of it can be attributed to individual neurons. We develop unsupervised methods for discovering important neurons in NMT models. Our methods rely on the intuition that different models learn similar properties, and do not require any costly external supervision. We show experimentally that translation quality depends on the discovered neurons, and find that many of them capture common linguistic phenomena. Finally, we show how to control NMT translations in predictable ways, by modifying activations of individual neurons.

## signSGD via Zeroth-Order Oracle

- Sijia Liu, Pin-Yu Chen, Xiangyi Chen, Mingyi Hong
- abstract@open-review(Poster): In this paper, we design and analyze a new zeroth-order (ZO) stochastic optimization algorithm, ZO-signSGD, which enjoys dual advantages of gradient-free operations and signSGD. The latter requires only the sign information of gradient estimates but is able to achieve a comparable or even better convergence speed than SGD-type algorithms. Our study shows that ZO signSGD requires $\sqrt{d}$ times more iterations than signSGD, leading to a convergence rate of $O(\sqrt{d}/\sqrt{T})$ under mild conditions, where $d$ is the number of optimization variables, and $T$ is the number of iterations. In addition, we analyze the effects of different types of gradient estimators on the convergence of ZO-signSGD, and propose two variants of ZO-signSGD that at least achieve $O(\sqrt{d}/\sqrt{T})$ convergence rate. On the application side we explore the connection between ZO-signSGD and black-box adversarial attacks in robust deep learning. Our empirical evaluations on image classification datasets MNIST and CIFAR-10 demonstrate the superior performance of ZO-signSGD on the generation of adversarial examples from black-box neural networks.

## DELTA: DEEP LEARNING TRANSFER USING FEATURE MAP WITH ATTENTION FOR CONVOLUTIONAL NETWORKS

- Xingjian Li, Haoyi Xiong, Hanchao Wang, Yuxuan Rao, Liping Liu, Jun Huan
- abstract@open-review(Poster): Transfer learning through fine-tuning a pre-trained neural network with an extremely large dataset, such as ImageNet, can significantly accelerate training while the accuracy is frequently bottlenecked by the limited dataset size of the new target task. To solve the problem, some regularization methods, constraining the outer layer weights of the target network using the starting point as references (SPAR), have been studied. In this paper, we propose a novel regularized transfer learning framework DELTA, namely DEep Learning Transfer using Feature Map with Attention. Instead of constraining the weights of neural network, DELTA aims to preserve the outer layer outputs of the target network. Specifically, in addition to minimizing the empirical loss, DELTA intends to align the outer layer outputs of two networks, through constraining a subset of feature maps that are precisely selected by attention that has been learned in an supervised learning manner. We evaluate DELTA with the state-of-the-art algorithms, including L2 and L2-SP. The experiment results show that our proposed method outperforms these baselines with higher accuracy for new tasks.

## Learning to Remember More with Less Memorization

- Hung Le, Truyen Tran, Svetha Venkatesh
- abstract@open-review(Oral): Memory-augmented neural networks consisting of a neural controller and an external memory have shown potentials in long-term sequential learning. Current RAM-like memory models maintain memory accessing every timesteps, thus they do not effectively leverage the short-term memory held in the controller. We hypothesize that this scheme of writing is suboptimal in memory utilization and introduces redundant computation. To validate our hypothesis, we derive a theoretical bound on the amount of information stored in a RAM-like system and formulate an optimization problem that maximizes the bound. The proposed solution dubbed Uniform Writing is proved to be optimal under the assumption of equal timestep contributions. To relax this assumption, we introduce modifications to the original solution, resulting in a solution termed Cached Uniform Writing. This method aims to balance between maximizing memorization and forgetting via overwriting mechanisms. Through an extensive set of experiments, we empirically demonstrate the advantages of our solutions over other recurrent architectures, claiming the state-of-the-arts in various sequential modeling tasks.

## Variance Networks: When Expectation Does Not Meet Your Expectations

- Kirill Neklyudov, Dmitry Molchanov, Arsenii Ashukha, Dmitry Vetrov
- abstract@open-review(Poster): Ordinary stochastic neural networks mostly rely on the expected values of their weights to make predictions, whereas the induced noise is mostly used to capture the uncertainty, prevent overfitting and slightly boost the performance through test-time averaging. In this paper, we introduce variance layers, a different kind of stochastic layers. Each weight of a variance layer follows a zero-mean distribution and is only parameterized by its variance. It means that each object is represented by a zero-mean distribution in the space of the activations. We show that such layers can learn surprisingly well, can serve as an efficient exploration tool in reinforcement learning tasks and provide a decent defense against adversarial attacks. We also show that a number of conventional Bayesian neural networks naturally converge to such zero-mean posteriors. We observe that in these cases such zero-mean parameterization leads to a much better training objective than more flexible conventional parameterizations where the mean is being learned.

## Deterministic Variational Inference for Robust Bayesian Neural Networks

- Anqi Wu, Sebastian Nowozin, Edward Meeds, Richard E. Turner, José Miguel Hernández-Lobato, Alexander L. Gaunt
- abstract@open-review(Oral): Bayesian neural networks (BNNs) hold great promise as a flexible and principled solution to deal with uncertainty when learning from finite data. Among approaches to realize probabilistic inference in deep neural networks, variational Bayes (VB) is theoretically grounded, generally applicable, and computationally efficient. With wide recognition of potential advantages, why is it that variational Bayes has seen very limited practical use for BNNs in real applications? We argue that variational inference in neural networks is fragile: successful implementations require careful initialization and tuning of prior variances, as well as controlling the variance of Monte Carlo gradient estimates. We provide two innovations that aim to turn VB into a robust inference tool for Bayesian neural networks: first, we introduce a novel deterministic method to approximate moments in neural networks, eliminating gradient variance; second, we introduce a hierarchical prior for parameters and a novel Empirical Bayes procedure for automatically selecting prior variances. Combining these two innovations, the resulting method is highly efficient and robust. On the application of heteroscedastic regression we demonstrate good predictive performance over alternative approaches.

## Conditional Network Embeddings

- Bo Kang, Jefrey Lijffijt, Tijl De Bie
- abstract@open-review(Poster): Network Embeddings (NEs) map the nodes of a given network into $d$-dimensional Euclidean space $\mathbb{R}^d$. Ideally, this mapping is such that 'similar' nodes are mapped onto nearby points, such that the NE can be used for purposes such as link prediction (if 'similar' means being 'more likely to be connected') or classification (if 'similar' means 'being more likely to have the same label'). In recent years various methods for NE have been introduced, all following a similar strategy: defining a notion of similarity between nodes (typically some distance measure within the network), a distance measure in the embedding space, and a loss function that penalizes large distances for similar nodes and small distances for dissimilar nodes.

A difficulty faced by existing methods is that certain networks are fundamentally hard to embed due to their structural properties: (approximate) multipartiteness, certain degree distributions, assortativity, etc. To overcome this, we introduce a conceptual innovation to the NE literature and propose to create \emph{Conditional Network Embeddings} (CNEs); embeddings that maximally add information with respect to given structural properties (e.g. node degrees, block densities, etc.). We use a simple Bayesian approach to achieve this, and propose a block stochastic gradient descent algorithm for fitting it efficiently.

We demonstrate that CNEs are superior for link prediction and multi-label classification when compared to state-of-the-art methods, and this without adding significant mathematical or computational complexity. Finally, we illustrate the potential of CNE for network visualization.

## Distribution-Interpolation Trade off in Generative Models

- Damian Leśniak, Igor Sieradzki, Igor Podolak
- abstract@open-review(Poster): We investigate the properties of multidimensional probability distributions in the context of latent space prior distributions of implicit generative models. Our work revolves around the phenomena arising while decoding linear interpolations between two random latent vectors -- regions of latent space in close proximity to the origin of the space are oversampled, which restricts the usability of linear interpolations as a tool to analyse the latent space. We show that the distribution mismatch can be eliminated completely by a proper choice of the latent probability distribution or using non-linear interpolations. We prove that there is a trade off between the interpolation being linear, and the latent distribution having even the most basic properties required for stable training, such as finite mean. We use the multidimensional Cauchy distribution as an example of the prior distribution, and also provide a general method of creating non-linear interpolations, that is easily applicable to a large family of commonly used latent distributions.

## Sample Efficient Adaptive Text-to-Speech

- Yutian Chen, Yannis Assael, Brendan Shillingford, David Budden, Scott Reed, Heiga Zen, Quan Wang, Luis C. Cobo, Andrew Trask, Ben Laurie, Caglar Gulcehre, Aäron van den Oord, Oriol Vinyals, Nando de Freitas
- abstract@open-review(Poster): We present a meta-learning approach for adaptive text-to-speech (TTS) with few data. During training, we learn a multi-speaker model using a shared conditional WaveNet core and independent learned embeddings for each speaker. The aim of training is not to produce a neural network with fixed weights, which is then deployed as a TTS system. Instead, the aim is to produce a network that requires few data at deployment time to rapidly adapt to new speakers. We introduce and benchmark three strategies: (i) learning the speaker embedding while keeping the WaveNet core fixed, (ii) fine-tuning the entire architecture with stochastic gradient descent, and (iii) predicting the speaker embedding with a trained neural network encoder. The experiments show that these approaches are successful at adapting the multi-speaker neural network to new speakers, obtaining state-of-the-art results in both sample naturalness and voice similarity with merely a few minutes of audio data from new speakers.

## Maximal Divergence Sequential Autoencoder for Binary Software Vulnerability Detection

- Tue Le, Tuan Nguyen, Trung Le, Dinh Phung, Paul Montague, Olivier De Vel, Lizhen Qu
- abstract@open-review(Poster): Due to the sharp increase in the severity of the threat imposed by software vulnerabilities, the detection of vulnerabilities in binary code has become an important concern in the software industry, such as the embedded systems industry, and in the field of computer security.

However, most of the work in binary code vulnerability detection has relied on handcrafted features which are manually chosen by a select few, knowledgeable domain experts. In this paper, we attempt to alleviate this severe binary vulnerability detection bottleneck by leveraging recent advances in deep learning representations and propose the Maximal Divergence Sequential Auto-Encoder. In particular, latent codes representing vulnerable and non-vulnerable binaries are encouraged to be maximally divergent, while still being able to maintain crucial information from the original binaries. We conducted extensive experiments to compare and contrast our proposed methods with the baselines, and the results show that our proposed methods outperform the baselines in all performance measures of interest.

## Sample Efficient Imitation Learning for Continuous Control

- Fumihiro Sasaki, Tetsuya Yohira, Atsuo Kawaguchi
- abstract@open-review(Poster): The goal of imitation learning (IL) is to enable a learner to imitate expert behavior given expert demonstrations. Recently, generative adversarial imitation learning (GAIL) has shown significant progress on IL for complex continuous tasks. However, GAIL and its extensions require a large number of environment interactions during training. In real-world environments, the more an IL method requires the learner to interact with the environment for better imitation, the more training time it requires, and the more damage it causes to the environments and the learner itself. We believe that IL algorithms could be more applicable to real-world problems if the number of interactions could be reduced. In this paper, we propose a model-free IL algorithm for continuous control. Our algorithm is made up mainly three changes to the existing adversarial imitation learning (AIL) methods – (a) adopting off-policy actor-critic (Off-PAC) algorithm to optimize the learner policy, (b) estimating the state-action value using off-policy samples without learning reward functions, and (c) representing the stochastic policy function so that its outputs are bounded. Experimental results show that our algorithm achieves competitive results with GAIL while significantly reducing the environment interactions.

## Practical lossless compression with latent variables using bits back coding

- James Townsend, Thomas Bird, David Barber
- abstract@open-review(Poster): Deep latent variable models have seen recent success in many data domains. Lossless compression is an application of these models which, despite having the potential to be highly useful, has yet to be implemented in a practical manner. We present `Bits Back with ANS' (BB-ANS), a scheme to perform lossless compression with latent variable models at a near optimal rate. We demonstrate this scheme by using it to compress the MNIST dataset with a variational auto-encoder model (VAE), achieving compression rates superior to standard methods with only a simple VAE. Given that the scheme is highly amenable to parallelization, we conclude that with a sufficiently high quality generative model this scheme could be used to achieve substantial improvements in compression rate with acceptable running time. We make our implementation available open source at https://github.com/bits-back/bits-back .

## Context-adaptive Entropy Model for End-to-end Optimized Image Compression

- Jooyoung Lee, Seunghyun Cho, Seung-Kwon Beack
- abstract@open-review(Poster): We propose a context-adaptive entropy model for use in end-to-end optimized image compression. Our model exploits two types of contexts, bit-consuming contexts and bit-free contexts, distinguished based upon whether additional bit allocation is required. Based on these contexts, we allow the model to more accurately estimate the distribution of each latent representation with a more generalized form of the approximation models, which accordingly leads to an enhanced compression performance. Based on the experimental results, the proposed method outperforms the traditional image codecs, such as BPG and JPEG2000, as well as other previous artificial-neural-network (ANN) based approaches, in terms of the peak signal-to-noise ratio (PSNR) and multi-scale structural similarity (MS-SSIM) index. The test code is publicly available at https://github.com/JooyoungLeeETRI/CA_Entropy_Model.

## Analysis of Quantized Models

- Lu Hou, Ruiliang Zhang, James T. Kwok
- abstract@open-review(Poster): Deep neural networks are usually huge, which significantly limits the deployment on low-end devices. In recent years, many weight-quantized models have been proposed. They have small storage and fast inference, but training can still be time-consuming. This can be improved with distributed learning. To reduce the high communication cost due to worker-server synchronization, recently gradient quantization has also been proposed to train deep networks with full-precision weights. In this paper, we theoretically study how the combination of both weight and gradient quantization affects convergence. We show that (i) weight-quantized models converge to an error related to the weight quantization resolution and weight dimension; (ii) quantizing gradients slows convergence by a factor related to the gradient quantization resolution and dimension; and (iii) clipping the gradient before quantization renders this factor dimension-free, thus allowing the use of fewer bits for gradient quantization. Empirical experiments confirm the theoretical convergence results, and demonstrate that quantized networks can speed up training and have comparable performance as full-precision networks.

## Generating Multiple Objects at Spatially Distinct Locations

- Tobias Hinz, Stefan Heinrich, Stefan Wermter
- abstract@open-review(Poster): Recent improvements to Generative Adversarial Networks (GANs) have made it possible to generate realistic images in high resolution based on natural language descriptions such as image captions. Furthermore, conditional GANs allow us to control the image generation process through labels or even natural language descriptions. However, fine-grained control of the image layout, i.e. where in the image specific objects should be located, is still difficult to achieve. This is especially true for images that should contain multiple distinct objects at different spatial locations. We introduce a new approach which allows us to control the location of arbitrarily many objects within an image by adding an object pathway to both the generator and the discriminator. Our approach does not need a detailed semantic layout but only bounding boxes and the respective labels of the desired objects are needed. The object pathway focuses solely on the individual objects and is iteratively applied at the locations specified by the bounding boxes. The global pathway focuses on the image background and the general image layout. We perform experiments on the Multi-MNIST, CLEVR, and the more complex MS-COCO data set. Our experiments show that through the use of the object pathway we can control object locations within images and can model complex scenes with multiple objects at various locations. We further show that the object pathway focuses on the individual objects and learns features relevant for these, while the global pathway focuses on global image characteristics and the image background.

## ANYTIME MINIBATCH: EXPLOITING STRAGGLERS IN ONLINE DISTRIBUTED OPTIMIZATION

- Nuwan Ferdinand, Haider Al-Lawati, Stark Draper, Matthew Nokleby
- abstract@open-review(Poster): Distributed optimization is vital in solving large-scale machine learning problems. A widely-shared feature of distributed optimization techniques is the requirement that all nodes complete their assigned tasks in each computational epoch before the system can proceed to the next epoch. In such settings, slow nodes, called stragglers, can greatly slow progress. To mitigate the impact of stragglers, we propose an online distributed optimization method called Anytime Minibatch. In this approach, all nodes are given a fixed time to compute the gradients of as many data samples as possible. The result is a variable per-node minibatch size. Workers then get a fixed communication time to average their minibatch gradients via several rounds of consensus, which are then used to update primal variables via dual averaging. Anytime Minibatch prevents stragglers from holding up the system without wasting the work that stragglers can complete. We present a convergence analysis and analyze the wall time performance. Our numerical results show that our approach is up to 1.5 times faster in Amazon EC2 and it is up to five times faster when there is greater variability in compute node performance.

## A rotation-equivariant convolutional neural network model of primary visual cortex

- Alexander S. Ecker, Fabian H. Sinz, Emmanouil Froudarakis, Paul G. Fahey, Santiago A. Cadena, Edgar Y. Walker, Erick Cobos, Jacob Reimer, Andreas S. Tolias, Matthias Bethge
- abstract@open-review(Poster): Classical models describe primary visual cortex (V1) as a filter bank of orientation-selective linear-nonlinear (LN) or energy models, but these models fail to predict neural responses to natural stimuli accurately. Recent work shows that convolutional neural networks (CNNs) can be trained to predict V1 activity more accurately, but it remains unclear which features are extracted by V1 neurons beyond orientation selectivity and phase invariance. Here we work towards systematically studying V1 computations by categorizing neurons into groups that perform similar computations. We present a framework for identifying common features independent of individual neurons' orientation selectivity by using a rotation-equivariant convolutional neural network, which automatically extracts every feature at multiple different orientations. We fit this rotation-equivariant CNN to responses of a population of 6000 neurons to natural images recorded in mouse primary visual cortex using two-photon imaging. We show that our rotation-equivariant network outperforms a regular CNN with the same number of feature maps and reveals a number of common features, which are shared by many V1 neurons and are pooled sparsely to predict neural activity. Our findings are a first step towards a powerful new tool to study the nonlinear functional organization of visual cortex.

## An analytic theory of generalization dynamics and transfer learning in deep linear networks

- Andrew K. Lampinen, Surya Ganguli
- abstract@open-review(Poster): Much attention has been devoted recently to the generalization puzzle in deep learning: large, deep networks can generalize well, but existing theories bounding generalization error are exceedingly loose, and thus cannot explain this striking performance. Furthermore, a major hope is that knowledge may transfer across tasks, so that multi-task learning can improve generalization on individual tasks. However we lack analytic theories that can quantitatively predict how the degree of knowledge transfer depends on the relationship between the tasks. We develop an analytic theory of the nonlinear dynamics of generalization in deep linear networks, both within and across tasks. In particular, our theory provides analytic solutions to the training and testing error of deep networks as a function of training time, number of examples, network size and initialization, and the task structure and SNR. Our theory reveals that deep networks progressively learn the most important task structure first, so that generalization error at the early stopping time primarily depends on task structure and is independent of network size. This suggests any tight bound on generalization error must take into account task structure, and explains observations about real data being learned faster than random data. Intriguingly our theory also reveals the existence of a learning algorithm that proveably out-performs neural network training through gradient descent. Finally, for transfer learning, our theory reveals that knowledge transfer depends sensitively, but computably, on the SNRs and input feature alignments of pairs of tasks.

## Are adversarial examples inevitable?

- Ali Shafahi, W. Ronny Huang, Christoph Studer, Soheil Feizi, Tom Goldstein
- abstract@open-review(Poster): A wide range of defenses have been proposed to harden neural networks against adversarial attacks. However, a pattern has emerged in which the majority of adversarial defenses are quickly broken by new attacks. Given the lack of success at generating robust defenses, we are led to ask a fundamental question: Are adversarial attacks inevitable? This paper analyzes adversarial examples from a theoretical perspective, and identifies fundamental bounds on the susceptibility of a classifier to adversarial attacks. We show that, for certain classes of problems, adversarial examples are inescapable. Using experiments, we explore the implications of theoretical guarantees for real-world problems and discuss how factors such as dimensionality and image complexity limit a classifier's robustness against adversarial examples.

## Directed-Info GAIL: Learning Hierarchical Policies from Unsegmented Demonstrations using Directed Information

- Mohit Sharma, Arjun Sharma, Nicholas Rhinehart, Kris M. Kitani
- abstract@open-review(Poster): The use of imitation learning to learn a single policy for a complex task that has multiple modes or hierarchical structure can be challenging. In fact, previous work has shown that when the modes are known, learning separate policies for each mode or sub-task can greatly improve the performance of imitation learning. In this work, we discover the interaction between sub-tasks from their resulting state-action trajectory sequences using a directed graphical model. We propose a new algorithm based on the generative adversarial imitation learning framework which automatically learns sub-task policies from unsegmented demonstrations. Our approach maximizes the directed information flow in the graphical model between sub-task latent variables and their generated trajectories. We also show how our approach connects with the existing Options framework, which is commonly used to learn hierarchical policies.

## M^3RL: Mind-aware Multi-agent Management Reinforcement Learning

- Tianmin Shu, Yuandong Tian
- abstract@open-review(Poster): Most of the prior work on multi-agent reinforcement learning (MARL) achieves optimal collaboration by directly learning a policy for each agent to maximize a common reward. In this paper, we aim to address this from a different angle. In particular, we consider scenarios where there are self-interested agents (i.e., worker agents) which have their own minds (preferences, intentions, skills, etc.) and can not be dictated to perform tasks they do not want to do. For achieving optimal coordination among these agents, we train a super agent (i.e., the manager) to manage them by first inferring their minds based on both current and past observations and then initiating contracts to assign suitable tasks to workers and promise to reward them with corresponding bonuses so that they will agree to work together. The objective of the manager is to maximize the overall productivity as well as minimize payments made to the workers for ad-hoc worker teaming. To train the manager, we propose Mind-aware Multi-agent Management Reinforcement Learning (M^3RL), which consists of agent modeling and policy learning. We have evaluated our approach in two environments, Resource Collection and Crafting, to simulate multi-agent management problems with various task settings and multiple designs for the worker agents. The experimental results have validated the effectiveness of our approach in modeling worker agents' minds online, and in achieving optimal ad-hoc teaming with good generalization and fast adaptation.

## Differentiable Learning-to-Normalize via Switchable Normalization

- Ping Luo, Jiamin Ren, Zhanglin Peng, Ruimao Zhang, Jingyu Li
- abstract@open-review(Poster): We address a learning-to-normalize problem by proposing Switchable Normalization (SN), which learns to select different normalizers for different normalization layers of a deep neural network. SN employs three distinct scopes to compute statistics (means and variances) including a channel, a layer, and a minibatch. SN switches between them by learning their importance weights in an end-to-end manner. It has several good properties. First, it adapts to various network architectures and tasks (see Fig.1). Second, it is robust to a wide range of batch sizes, maintaining high performance even when small minibatch is presented (e.g. 2 images/GPU). Third, SN does not have sensitive hyper-parameter, unlike group normalization that searches the number of groups as a hyper-parameter. Without bells and whistles, SN outperforms its counterparts on various challenging benchmarks, such as ImageNet, COCO, CityScapes, ADE20K, and Kinetics. Analyses of SN are also presented. We hope SN will help ease the usage and understand the normalization techniques in deep learning. The code of SN will be released.

## Supervised Policy Update for Deep Reinforcement Learning

- Quan Vuong, Yiming Zhang, Keith W. Ross

- abstract@open-review(Poster): We propose a new sample-efficient methodology, called Supervised Policy Update (SPU), for deep reinforcement learning. Starting with data generated by the current policy, SPU formulates and solves a constrained optimization problem in the non-parameterized proximal policy space. Using supervised regression, it then converts the optimal non-parameterized policy to a parameterized policy, from which it draws new samples. The methodology is general in that it applies to both discrete and continuous action spaces, and can handle a wide variety of proximity constraints for the non-parameterized optimization problem. We show how the Natural Policy Gradient and Trust Region Policy Optimization (NPG/TRPO) problems, and the Proximal Policy Optimization (PPO) problem can be addressed by this methodology. The SPU implementation is much simpler than TRPO. In terms of sample efficiency, our extensive experiments show SPU outperforms TRPO in Mujoco simulated robotic tasks and outperforms PPO in Atari video game tasks.

## Adversarial Domain Adaptation for Stable Brain-Machine Interfaces

- Ali Farshchian, Juan A. Gallego, Joseph P. Cohen, Yoshua Bengio, Lee E. Miller, Sara A. Solla
- abstract@open-review(Poster): Brain-Machine Interfaces (BMIs) have recently emerged as a clinically viable option to restore voluntary movements after paralysis. These devices are based on the ability to extract information about movement intent from neural signals recorded using multi-electrode arrays chronically implanted in the motor cortices of the brain. However, the inherent loss and turnover of recorded neurons requires repeated recalibrations of the interface, which can potentially alter the day-to-day user experience. The resulting need for continued user adaptation interferes with the natural, subconscious use of the BMI. Here, we introduce a new computational approach that decodes movement intent from a low-dimensional latent representation of the neural data. We implement various domain adaptation methods to stabilize the interface over significantly long times. This includes Canonical Correlation Analysis used to align the latent variables across days; this method requires prior point-to-point correspondence of the time series across domains. Alternatively, we match the empirical probability distributions of the latent variables across days through the minimization of their Kullback-Leibler divergence. These two methods provide a significant and comparable improvement in the performance of the interface. However, implementation of an Adversarial Domain Adaptation Network trained to match the empirical probability distribution of the residuals of the reconstructed neural signals outperforms the two methods based on latent variables, while requiring remarkably few data points to solve the domain adaptation problem.

## LEARNING FACTORIZED REPRESENTATIONS FOR OPEN-SET DOMAIN ADAPTATION

- Mahsa Baktashmotlagh, Masoud Faraki, Tom Drummond, Mathieu Salzmann
- abstract@open-review(Poster): Domain adaptation for visual recognition has undergone great progress in the past few years. Nevertheless, most existing methods work in the so-called closed-set scenario, assuming that the classes depicted by the target images are exactly the same as those of the source domain. In this paper, we tackle the more challenging, yet more realistic case of open-set domain adaptation, where new, unknown classes can be present in the target data. While, in the unsupervised scenario, one cannot expect to be able to identify each specific new class, we aim to automatically detect which samples belong to these new classes and discard them from the recognition process. To this end, we rely on the intuition that the source and target samples depicting the known classes can be generated by a shared subspace, whereas the target samples from unknown classes come from a different, private subspace. We therefore introduce a framework that factorizes the data into shared and private parts, while encouraging the shared representation to be discriminative. Our experiments on standard benchmarks evidence that our approach significantly outperforms the state-of-the-art in open-set domain adaptation.

## Slalom: Fast, Verifiable and Private Execution of Neural Networks in Trusted Hardware

- Florian Tramer, Dan Boneh
- abstract@open-review(Oral): As Machine Learning (ML) gets applied to security-critical or sensitive domains, there is a growing need for integrity and privacy for outsourced ML computations. A pragmatic solution comes from Trusted Execution Environments (TEEs), which use hardware and software protections to isolate sensitive computations from the untrusted software stack. However, these isolation guarantees come at a price in performance, compared to untrusted alternatives. This paper initiates the study of high performance execution of Deep Neural Networks (DNNs) in TEEs by efficiently partitioning DNN computations between trusted and untrusted devices. Building upon an efficient outsourcing scheme for matrix multiplication, we propose Slalom, a framework that securely delegates execution of all linear layers in a DNN from a TEE (e.g., Intel SGX or Sanctum) to a faster, yet untrusted, co-located processor. We evaluate Slalom by running DNNs in an Intel SGX enclave, which selectively delegates work to an untrusted GPU. For canonical DNNs (VGG16, MobileNet and ResNet variants) we obtain 6x to 20x increases in throughput for verifiable inference, and 4x to 11x for verifiable and private inference.

## Learning to Understand Goal Specifications by Modelling Reward

- Dzmitry Bahdanau, Felix Hill, Jan Leike, Edward Hughes, Arian Hosseini, Pushmeet Kohli, Edward Grefenstette
- abstract@open-review(Poster): Recent work has shown that deep reinforcement-learning agents can learn to follow language-like instructions from infrequent environment rewards. However, this places on environment designers the onus of designing language-conditional reward functions which may not be easily or tractably implemented as the complexity of the environment and the language scales. To overcome this limitation, we present a framework within which instruction-conditional RL agents are trained using rewards obtained not from the environment, but from reward models which are jointly trained from expert examples. As reward models improve, they learn to accurately reward agents for completing tasks for environment configurations---and for instructions---not present amongst the expert data. This framework effectively separates the representation of what instructions require from how they can be executed. In a simple grid world, it enables an agent to learn a range of commands requiring interaction with blocks and understanding of spatial relations and underspecified abstract arrangements. We further show the method allows our agent to adapt to changes in the environment without requiring new expert examples.

## Dynamic Sparse Graph for Efficient Deep Learning

- Liu Liu, Lei Deng, Xing Hu, Maohua Zhu, Guoqi Li, Yufei Ding, Yuan Xie
- abstract@open-review(Poster): We propose to execute deep neural networks (DNNs) with dynamic and sparse graph (DSG) structure for compressive memory and accelerative execution during both training and inference. The great success of DNNs motivates the pursuing of lightweight models for the deployment onto embedded devices. However, most of the previous studies optimize for inference while neglect training or even complicate it. Training is far more intractable, since (i) the neurons dominate the memory cost rather than the weights in inference; (ii) the dynamic activation makes previous sparse acceleration via one-off optimization on fixed weight invalid; (iii) batch normalization (BN) is critical for maintaining accuracy while its activation reorganization damages the sparsity. To address these issues, DSG activates only a small amount of neurons with high selectivity at each iteration via a dimensionreduction search and obtains the BN compatibility via a double-mask selection. Experiments show significant memory saving (1.7-4.5x) and operation reduction (2.3-4.4x) with little accuracy loss on various benchmarks.

## Hierarchical interpretations for neural network predictions

- Chandan Singh, W. James Murdoch, Bin Yu
- abstract@open-review(Poster): Deep neural networks (DNNs) have achieved impressive predictive performance due to their ability to learn complex, non-linear relationships between variables. However, the inability to effectively visualize these relationships has led to DNNs being characterized as black boxes and consequently limited their applications. To ameliorate this problem, we introduce the use of hierarchical interpretations to explain DNN predictions through our proposed method: agglomerative contextual decomposition (ACD). Given a prediction from a trained DNN, ACD produces a

hierarchical clustering of the input features, along with the contribution of each cluster to the final prediction. This hierarchy is optimized to identify clusters of features that the DNN learned are predictive. We introduce ACD using examples from Stanford Sentiment Treebank and ImageNet, in order to diagnose incorrect predictions, identify dataset bias, and extract polarizing phrases of varying lengths. Through human experiments, we demonstrate that ACD enables users both to identify the more accurate of two DNNs and to better trust a DNN's outputs. We also find that ACD's hierarchy is largely robust to adversarial perturbations, implying that it captures fundamental aspects of the input and ignores spurious noise.

## Post Selection Inference with Incomplete Maximum Mean Discrepancy Estimator

- Makoto Yamada, Denny Wu, Yao-Hung Hubert Tsai, Hirofumi Ohta, Ruslan Salakhutdinov, Ichiro Takeuchi, Kenji Fukumizu
- abstract@open-review(Poster): Measuring divergence between two distributions is essential in machine learning and statistics and has various applications including binary classification, change point detection, and two-sample test. Furthermore, in the era of big data, designing divergence measure that is interpretable and can handle high-dimensional and complex data becomes extremely important. In this paper, we propose a post selection inference (PSI) framework for divergence measure, which can select a set of statistically significant features that discriminate two distributions. Specifically, we employ an additive variant of maximum mean discrepancy (MMD) for features and introduce a general hypothesis test for PSI. A novel MMD estimator using the incomplete U-statistics, which has an asymptotically normal distribution (under mild assumptions) and gives high detection power in PSI, is also proposed and analyzed theoretically. Through synthetic and real-world feature selection experiments, we show that the proposed framework can successfully detect statistically significant features. Last, we propose a sample selection framework for analyzing different members in the Generative Adversarial Networks (GANs) family.

## Diffusion Scattering Transforms on Graphs

- Fernando Gama, Alejandro Ribeiro, Joan Bruna
- abstract@open-review(Poster): Stability is a key aspect of data analysis. In many applications, the natural notion of stability is geometric, as illustrated for example in computer vision. Scattering transforms construct deep convolutional representations which are certified stable to input deformations. This stability to deformations can be interpreted as stability with respect to changes in the metric structure of the domain.

In this work, we show that scattering transforms can be generalized to non-Euclidean domains using diffusion wavelets, while preserving a notion of stability with respect to metric changes in the domain, measured with diffusion maps. The resulting representation is stable to metric perturbations of the domain while being able to capture "high-frequency" information, akin to the Euclidean Scattering.

## Preconditioner on Matrix Lie Group for SGD

- Xi-Lin Li
- abstract@open-review(Poster): We study two types of preconditioners and preconditioned stochastic gradient descent (SGD) methods in a unified framework. We call the first one the Newton type due to its close relationship to the Newton method, and the second one the Fisher type as its preconditioner is closely related to the inverse of Fisher information matrix. Both preconditioners can be derived from one framework, and efficiently estimated on any matrix Lie groups designated by the user using natural or relative gradient descent minimizing certain preconditioner estimation criteria. Many existing preconditioners and methods, e.g., RMSProp, Adam, KFAC, equilibrated SGD, batch normalization, etc., are special cases of or closely related to either the Newton type or the Fisher type ones. Experimental results on relatively large scale machine learning problems are reported for performance study.

## DPSNet: End-to-end Deep Plane Sweep Stereo

- Sunghoon Im, Hae-Gon Jeon, Stephen Lin, In So Kweon
- abstract@open-review(Poster): Multiview stereo aims to reconstruct scene depth from images acquired by a camera under arbitrary motion. Recent methods address this problem through deep learning, which can utilize semantic cues to deal with challenges such as textureless and reflective regions. In this paper, we present a convolutional neural network called DPSNet (Deep Plane Sweep Network) whose design is inspired by best practices of traditional geometry-based approaches. Rather than directly estimating depth and/or optical flow correspondence from image pairs as done in many previous deep learning methods, DPSNet takes a plane sweep approach that involves building a cost volume from deep features using the plane sweep algorithm, regularizing the cost volume via a context-aware cost aggregation, and regressing the depth map from the cost volume. The cost volume is constructed using a differentiable warping process that allows for end-to-end training of the network. Through the effective incorporation of conventional multiview stereo concepts within a deep learning framework, DPSNet achieves state-of-the-art reconstruction results on a variety of challenging datasets.

## DARTS: Differentiable Architecture Search

- Hanxiao Liu, Karen Simonyan, Yiming Yang
- abstract@open-review(Poster): This paper addresses the scalability challenge of architecture search by formulating the task in a differentiable manner. Unlike conventional approaches of applying evolution or reinforcement learning over a discrete and non-differentiable search space, our method is based on the continuous relaxation of the architecture representation, allowing efficient search of the architecture using gradient descent. Extensive experiments on CIFAR-10, ImageNet, Penn Treebank and WikiText-2 show that our algorithm excels in discovering high-performance convolutional architectures for image classification and recurrent architectures for language modeling, while being orders of magnitude faster than state-of-the-art non-differentiable techniques.

## Adversarial Reprogramming of Neural Networks

- Gamaleldin F. Elsayed, Ian Goodfellow, Jascha Sohl-Dickstein
- abstract@open-review(Poster): Deep neural networks are susceptible to adversarial attacks. In computer vision, well-crafted perturbations to images can cause neural networks to make mistakes such as confusing a cat with a computer. Previous adversarial attacks have been designed to degrade performance of models or cause machine learning models to produce specific outputs chosen ahead of time by the attacker. We introduce attacks that instead reprogram the target model to perform a task chosen by the attacker without the attacker needing to specify or compute the desired output for each test-time input. This attack finds a single adversarial perturbation, that can be added to all test-time inputs to a machine learning model in order to cause the model to perform a task chosen by the adversary—even if the model was not trained to do this task. These perturbations can thus be considered a program for the new task. We demonstrate adversarial reprogramming on six ImageNet classification models, repurposing these models to perform a counting task, as well as classification tasks: classification of MNIST and CIFAR-10 examples presented as inputs to the ImageNet model.

## Opportunistic Learning: Budgeted Cost-Sensitive Learning from Data Streams

- Mohammad Kachuee, Orpaz Goldstein, Kimmo Kärkkäinen, Sajad Darabi, Majid Sarrafzadeh
- abstract@open-review(Poster): In many real-world learning scenarios, features are only acquirable at a cost constrained under a budget. In this paper, we propose a novel approach for cost-sensitive feature acquisition at the prediction-time. The suggested method acquires features incrementally based on a context-aware feature-value function. We formulate the problem in the reinforcement learning paradigm, and introduce a reward function based on the utility of each feature. Specifically, MC dropout sampling is used to measure expected variations of the model uncertainty which is used as a feature-value

function. Furthermore, we suggest sharing representations between the class predictor and value function estimator networks. The suggested approach is completely online and is readily applicable to stream learning setups. The solution is evaluated on three different datasets including the well-known MNIST dataset as a benchmark as well as two cost-sensitive datasets: Yahoo Learning to Rank and a dataset in the medical domain for diabetes classification. According to the results, the proposed method is able to efficiently acquire features and make accurate predictions.

## INVASE: Instance-wise Variable Selection using Neural Networks

- Jinsung Yoon, James Jordon, Mihaela van der Schaar
- abstract@open-review(Poster): The advent of big data brings with it data with more and more dimensions and thus a growing need to be able to efficiently select which features to use for a variety of problems. While global feature selection has been a well-studied problem for quite some time, only recently has the paradigm of instance-wise feature selection been developed. In this paper, we propose a new instance-wise feature selection method, which we term INVASE. INVASE consists of 3 neural networks, a selector network, a predictor network and a baseline network which are used to train the selector network using the actor-critic methodology. Using this methodology, INVASE is capable of flexibly discovering feature subsets of a different size for each instance, which is a key limitation of existing state-of-the-art methods. We demonstrate through a mixture of synthetic and real data experiments that INVASE significantly outperforms state-of-the-art benchmarks.

## The relativistic discriminator: a key element missing from standard GAN

- Alexia Jolicoeur-Martineau
- abstract@open-review(Poster): In standard generative adversarial network (SGAN), the discriminator estimates the probability that the input data is real. The generator is trained to increase the probability that fake data is real. We argue that it should also simultaneously decrease the probability that real data is real because 1) this would account for a priori knowledge that half of the data in the mini-batch is fake, 2) this would be observed with divergence minimization, and 3) in optimal settings, SGAN would be equivalent to integral probability metric (IPM) GANs.

We show that this property can be induced by using a relativistic discriminator which estimate the probability that the given real data is more realistic than a randomly sampled fake data. We also present a variant in which the discriminator estimate the probability that the given real data is more realistic than fake data, on average. We generalize both approaches to non-standard GAN loss functions and we refer to them respectively as Relativistic GANs (RGANs) and Relativistic average GANs (RaGANs). We show that IPM-based GANs are a subset of RGANs which use the identity function.

Empirically, we observe that 1) RGANs and RaGANs are significantly more stable and generate higher quality data samples than their non-relativistic counterparts, 2) Standard RaGAN with gradient penalty generate data of better quality than WGAN-GP while only requiring a single discriminator update per generator update (reducing the time taken for reaching the state-of-the-art by 400%), and 3) RaGANs are able to generate plausible high resolutions images (256x256) from a very small sample (N=2011), while GAN and LSGAN cannot; these images are of significantly better quality than the ones generated by WGAN-GP and SGAN with spectral normalization.

The code is freely available on https://github.com/AlexiaJM/RelativisticGAN.

## A Kernel Random Matrix-Based Approach for Sparse PCA

- Mohamed El Amine Seddik, Mohamed Tamaazousti, Romain Couillet
- abstract@open-review(Poster): In this paper, we present a random matrix approach to recover sparse principal components from n p-dimensional vectors. Specifically, considering the large dimensional setting where $n, p \to \infty$ with $p/n \to c \in (0, \infty)$ and under Gaussian vector observations, we study kernel random matrices of the type $f(\hat{C})$, where f is a three-times continuously differentiable function applied entry-wise to the sample covariance matrix $\hat{C}$ of the data. Then, assuming that the principal components are sparse, we show that taking f in such a way that $f'(0) = f''(0) = 0$ allows for powerful recovery of the principal components, thereby generalizing previous ideas involving more specific f functions such as the soft-thresholding function.

## Caveats for information bottleneck in deterministic scenarios

- Artemy Kolchinsky, Brendan D. Tracey, Steven Van Kuyk
- abstract@open-review(Poster): Information bottleneck (IB) is a method for extracting information from one random variable X that is relevant for predicting another random variable Y. To do so, IB identifies an intermediate "bottleneck" variable T that has low mutual information I(X;T) and high mutual information I(Y;T). The "IB curve" characterizes the set of bottleneck variables that achieve maximal I(Y;T) for a given I(X;T), and is typically explored by maximizing the "IB Lagrangian", I(Y;T) - βI(X;T). In some cases, Y is a deterministic function of X, including many classification problems in supervised learning where the output class Y is a deterministic function of the input X. We demonstrate three caveats when using IB in any situation where Y is a deterministic function of X: (1) the IB curve cannot be recovered by maximizing the IB Lagrangian for different values of β; (2) there are "uninteresting" trivial solutions at all points of the IB curve; and (3) for multi-layer classifiers that achieve low prediction error, different layers cannot exhibit a strict trade-off between compression and prediction, contrary to a recent proposal. We also show that when Y is a small perturbation away from being a deterministic function of X, these three caveats arise in an approximate way. To address problem (1), we propose a functional that, unlike the IB Lagrangian, can recover the IB curve in all cases. We demonstrate the three caveats on the MNIST dataset.

## Learning Localized Generative Models for 3D Point Clouds via Graph Convolution

- Diego Valsesia, Giulia Fracastoro, Enrico Magli
- abstract@open-review(Poster): Point clouds are an important type of geometric data and have widespread use in computer graphics and vision. However, learning representations for point clouds is particularly challenging due to their nature as being an unordered collection of points irregularly distributed in 3D space. Graph convolution, a generalization of the convolution operation for data defined over graphs, has been recently shown to be very successful at extracting localized features from point clouds in supervised or semi-supervised tasks such as classification or segmentation. This paper studies the unsupervised problem of a generative model exploiting graph convolution. We focus on the generator of a GAN and define methods for graph convolution when the graph is not known in advance as it is the very output of the generator. The proposed architecture learns to generate localized features that approximate graph embeddings of the output geometry. We also study the problem of defining an upsampling layer in the graph-convolutional generator, such that it learns to exploit a self-similarity prior on the data distribution to sample more effectively.

## Understanding and Improving Interpolation in Autoencoders via an Adversarial Regularizer

- David Berthelot, *Colin Raffel*, Aurko Roy, Ian Goodfellow
- abstract@open-review(Poster): Autoencoders provide a powerful framework for learning compressed representations by encoding all of the information needed to reconstruct a data point in a latent code. In some cases, autoencoders can "interpolate": By decoding the convex combination of the latent codes for two datapoints, the autoencoder can produce an output which semantically mixes characteristics from the datapoints. In this paper, we propose a regularization procedure which encourages interpolated outputs to appear more realistic by fooling a critic network which has been trained to recover the mixing coefficient from interpolated data. We then develop a simple benchmark task where we can quantitatively measure the extent to which various autoencoders can interpolate and show that our regularizer dramatically improves interpolation in this setting. We also demonstrate empirically that our

regularizer produces latent codes which are more effective on downstream tasks, suggesting a possible link between interpolation abilities and learning useful representations.

## Adversarial Imitation via Variational Inverse Reinforcement Learning

- Ahmed H. Qureshi, Byron Boots, Michael C. Yip
- abstract@open-review(Poster): We consider a problem of learning the reward and policy from expert examples under unknown dynamics. Our proposed method builds on the framework of generative adversarial networks and introduces the empowerment-regularized maximum-entropy inverse reinforcement learning to learn near-optimal rewards and policies. Empowerment-based regularization prevents the policy from overfitting to expert demonstrations, which advantageously leads to more generalized behaviors that result in learning near-optimal rewards. Our method simultaneously learns empowerment through variational information maximization along with the reward and policy under the adversarial learning formulation. We evaluate our approach on various high-dimensional complex control tasks. We also test our learned rewards in challenging transfer learning problems where training and testing environments are made to be different from each other in terms of dynamics or structure. The results show that our proposed method not only learns near-optimal rewards and policies that are matching expert behavior but also performs significantly better than state-of-the-art inverse reinforcement learning algorithms.

## Generating Liquid Simulations with Deformation-aware Neural Networks

- Lukas Prantl, Boris Bonev, Nils Thuerey
- abstract@open-review(Poster): We propose a novel approach for deformation-aware neural networks that learn the weighting and synthesis of dense volumetric deformation fields. Our method specifically targets the space-time representation of physical surfaces from liquid simulations. Liquids exhibit highly complex, non-linear behavior under changing simulation conditions such as different initial conditions. Our algorithm captures these complex phenomena in two stages: a first neural network computes a weighting function for a set of pre-computed deformations, while a second network directly generates a deformation field for refining the surface. Key for successful training runs in this setting is a suitable loss function that encodes the effect of the deformations, and a robust calculation of the corresponding gradients. To demonstrate the effectiveness of our approach, we showcase our method with several complex examples of flowing liquids with topology changes. Our representation makes it possible to rapidly generate the desired implicit surfaces. We have implemented a mobile application to demonstrate that real-time interactions with complex liquid effects are possible with our approach.

## L2-Nonexpansive Neural Networks

- Haifeng Qian, Mark N. Wegman
- abstract@open-review(Poster): This paper proposes a class of well-conditioned neural networks in which a unit amount of change in the inputs causes at most a unit amount of change in the outputs or any of the internal layers. We develop the known methodology of controlling Lipschitz constants to realize its full potential in maximizing robustness, with a new regularization scheme for linear layers, new ways to adapt nonlinearities and a new loss function. With MNIST and CIFAR-10 classifiers, we demonstrate a number of advantages. Without needing any adversarial training, the proposed classifiers exceed the state of the art in robustness against white-box L2-bounded adversarial attacks. They generalize better than ordinary networks from noisy data with partially random labels. Their outputs are quantitatively meaningful and indicate levels of confidence and generalization, among other desirable properties.

## Deep, Skinny Neural Networks are not Universal Approximators

- Jesse Johnson
- abstract@open-review(Poster): In order to choose a neural network architecture that will be effective for a particular modeling problem, one must understand the limitations imposed by each of the potential options. These limitations are typically described in terms of information theoretic bounds, or by comparing the relative complexity needed to approximate example functions between different architectures. In this paper, we examine the topological constraints that the architecture of a neural network imposes on the level sets of all the functions that it is able to approximate. This approach is novel for both the nature of the limitations and the fact that they are independent of network depth for a broad family of activation functions.

## Large Scale Graph Learning From Smooth Signals

- Vassilis Kalofolias, Nathanaël Perraudin
- abstract@open-review(Poster): Graphs are a prevalent tool in data science, as they model the inherent structure of the data. Typically they are constructed either by connecting nearest samples, or by learning them from data, solving an optimization problem. While graph learning does achieve a better quality, it also comes with a higher computational cost. In particular, the current state-of-the-art model cost is O(n^2) for n samples. In this paper, we show how to scale it, obtaining an approximation with leading cost of O(n log(n)), with quality that approaches the exact graph learning model. Our algorithm uses known approximate nearest neighbor techniques to reduce the number of variables, and automatically selects the correct parameters of the model, requiring a single intuitive input: the desired edge density.

## RotDCF: Decomposition of Convolutional Filters for Rotation-Equivariant Deep Networks

- Xiuyuan Cheng, Qiang Qiu, Robert Calderbank, Guillermo Sapiro
- abstract@open-review(Poster): Explicit encoding of group actions in deep features makes it possible for convolutional neural networks (CNNs) to handle global deformations of images, which is critical to success in many vision tasks. This paper proposes to decompose the convolutional filters over joint steerable bases across the space and the group geometry simultaneously, namely a rotation-equivariant CNN with decomposed convolutional filters (RotDCF). This decomposition facilitates computing the joint convolution, which is proved to be necessary for the group equivariance. It significantly reduces the model size and computational complexity while preserving performance, and truncation of the bases expansion serves implicitly to regularize the filters. On datasets involving in-plane and out-of-plane object rotations, RotDCF deep features demonstrate greater robustness and interpretability than regular CNNs. The stability of the equivariant representation to input variations is also proved theoretically. The RotDCF framework can be extended to groups other than rotations, providing a general approach which achieves both group equivariance and representation stability at a reduced model size.

## Per-Tensor Fixed-Point Quantization of the Back-Propagation Algorithm

- Charbel Sakr, Naresh Shanbhag
- abstract@open-review(Poster): The high computational and parameter complexity of neural networks makes their training very slow and difficult to deploy on energy and storage-constrained comput- ing systems. Many network complexity reduction techniques have been proposed including fixed-point implementation. However, a systematic approach for design- ing full fixed-point training and inference of deep neural networks remains elusive. We describe a precision assignment methodology for neural network training in which all network parameters, i.e., activations and weights in the feedforward path, gradients and weight accumulators in the feedback path, are assigned close to minimal precision. The precision assignment is derived analytically and enables tracking the convergence behavior of the full precision training, known to converge a priori. Thus, our work leads to a systematic methodology of determining suit- able precision for fixed-point training. The near optimality (minimality) of the resulting precision assignment is validated empirically for four networks on the CIFAR-10, CIFAR-100, and SVHN datasets. The complexity reduction arising from our approach is compared with other fixed-point neural network designs.

## Understanding Straight-Through Estimator in Training Activation Quantized Neural Nets

- Penghang Yin, Jiancheng Lyu, Shuai Zhang, Stanley Osher, Yingyong Qi, Jack Xin
- abstract@open-review(Poster): Training activation quantized neural networks involves minimizing a piecewise constant training loss whose gradient vanishes almost everywhere, which is undesirable for the standard back-propagation or chain rule. An empirical way around this issue is to use a straight-through estimator (STE) (Bengio et al., 2013) in the backward pass only, so that the "gradient" through the modified chain rule becomes non-trivial. Since this unusual "gradient" is certainly not the gradient of loss function, the following question arises: why searching in its negative direction minimizes the training loss? In this paper, we provide the theoretical justification of the concept of STE by answering this question. We consider the problem of learning a two-linear-layer network with binarized ReLU activation and Gaussian input data. We shall refer to the unusual "gradient" given by the STE-modifed chain rule as coarse gradient. The choice of STE is not unique. We prove that if the STE is properly chosen, the expected coarse gradient correlates positively with the population gradient (not available for the training), and its negation is a descent direction for minimizing the population loss. We further show the associated coarse gradient descent algorithm converges to a critical point of the population loss minimization problem. Moreover, we show that a poor choice of STE leads to instability of the training algorithm near certain local minima, which is verified with CIFAR-10 experiments.

## Convolutional Neural Networks on Non-uniform Geometrical Signals Using Euclidean Spectral Transformation

- Chiyu Max Jiang, Dequan Wang, Jingwei Huang, Philip Marcus, Matthias Niessner
- abstract@open-review(Poster): Convolutional Neural Networks (CNN) have been successful in processing data signals that are uniformly sampled in the spatial domain (e.g., images). However, most data signals do not natively exist on a grid, and in the process of being sampled onto a uniform physical grid suffer significant aliasing error and information loss. Moreover, signals can exist in different topological structures as, for example, points, lines, surfaces and volumes. It has been challenging to analyze signals with mixed topologies (for example, point cloud with surface mesh). To this end, we develop mathematical formulations for Non-Uniform Fourier Transforms (NUFT) to directly, and optimally, sample nonuniform data signals of different topologies defined on a simplex mesh into the spectral domain with no spatial sampling error. The spectral transform is performed in the Euclidean space, which removes the translation ambiguity from works on the graph spectrum. Our representation has four distinct advantages: (1) the process causes no spatial sampling error during initial sampling, (2) the generality of this approach provides a unified framework for using CNNs to analyze signals of mixed topologies, (3) it allows us to leverage state-of-the-art backbone CNN architectures for effective learning without having to design a particular architecture for a particular data structure in an ad-hoc fashion, and (4) the representation allows weighted meshes where each element has a different weight (i.e., texture) indicating local properties. We achieve good results on-par with state-of-the-art for 3D shape retrieval task, and new state-of-the-art for point cloud to surface reconstruction task.

## On Random Deep Weight-Tied Autoencoders: Exact Asymptotic Analysis, Phase Transitions, and Implications to Training

- Ping Li, Phan-Minh Nguyen
- abstract@open-review(Oral): We study the behavior of weight-tied multilayer vanilla autoencoders under the assumption of random weights. Via an exact characterization in the limit of large dimensions, our analysis reveals interesting phase transition phenomena when the depth becomes large. This, in particular, provides quantitative answers and insights to three questions that were yet fully understood in the literature. Firstly, we provide a precise answer on how the random deep weight-tied autoencoder model performs “approximate inference” as posed by Scellier et al. (2018), and its connection to reversibility considered by several theoretical studies. Secondly, we show that deep autoencoders display a higher degree of sensitivity to perturbations in the parameters, distinct from the shallow counterparts. Thirdly, we obtain insights on pitfalls in training initialization practice, and demonstrate experimentally that it is possible to train a deep autoencoder, even with the tanh activation and a depth as large as 200 layers, without resorting to techniques such as layer-wise pre-training or batch normalization. Our analysis is not specific to any depths or any Lipschitz activations, and our analytical techniques may have broader applicability.

## Training for Faster Adversarial Robustness Verification via Inducing ReLU Stability

- Kai Y. Xiao, Vincent Tjeng, Nur Muhammad (Mahi) Shafiullah, Aleksander Madry
- abstract@open-review(Poster): We explore the concept of co-design in the context of neural network verification. Specifically, we aim to train deep neural networks that not only are robust to adversarial perturbations but also whose robustness can be verified more easily. To this end, we identify two properties of network models - weight sparsity and so-called ReLU stability - that turn out to significantly impact the complexity of the corresponding verification task. We demonstrate that improving weight sparsity alone already enables us to turn computationally intractable verification problems into tractable ones. Then, improving ReLU stability leads to an additional 4-13x speedup in verification times. An important feature of our methodology is its "universality," in the sense that it can be used with a broad range of training procedures and verification approaches.

## Smoothing the Geometry of Probabilistic Box Embeddings

- Xiang Li, Luke Vilnis, Dongxu Zhang, Michael Boratko, Andrew McCallum
- abstract@open-review(Oral): There is growing interest in geometrically-inspired embeddings for learning hierarchies, partial orders, and lattice structures, with natural applications to transitive relational data such as entailment graphs. Recent work has extended these ideas beyond deterministic hierarchies to probabilistically calibrated models, which enable learning from uncertain supervision and inferring soft-inclusions among concepts, while maintaining the geometric inductive bias of hierarchical embedding models. We build on the Box Lattice model of Vilnis et al. (2018), which showed promising results in modeling soft-inclusions through an overlapping hierarchy of sets, parameterized as high-dimensional hyperrectangles (boxes). However, the hard edges of the boxes present difficulties for standard gradient based optimization; that work employed a special surrogate function for the disjoint case, but we find this method to be fragile. In this work, we present a novel hierarchical embedding model, inspired by a relaxation of box embeddings into parameterized density functions using Gaussian convolutions over the boxes. Our approach provides an alternative surrogate to the original lattice measure that improves the robustness of optimization in the disjoint case, while also preserving the desirable properties with respect to the original lattice. We demonstrate increased or matching performance on WordNet hypernymy prediction, Flickr caption entailment, and a MovieLens-based market basket dataset. We show especially marked improvements in the case of sparse data, where many conditional probabilities should be low, and thus boxes should be nearly disjoint.